Network Science Analytics
Option Applied Math and M.Sc. in DSBA

Lecture 2B
Power-law degree distribution and the Preferential Attachment model

Fragkiskos Malliaros

Friday, January 26, 2018

Acknowledgements

- The lecture is partially based on material by
  - Jure Leskovec, Stanford University
  - Aaron Clauset, CU Boulder
  - Manos Papagelis, York University
  - Gonzalo Mateos, University of Rochester
  - Albert-László Barabási, Northeastern University
  - Christos Faloutsos, CMU
  - Danai Koutra, University of Michigan
  - R. Zafarani, M. A. Abbasi, and H. Liu, Social Media Mining: An Introduction, Cambridge University Press, 2014. Free book and slides at http://socialmediamining.info/

Thank you!

Real network are not random

They follow properties that deviate from randomness

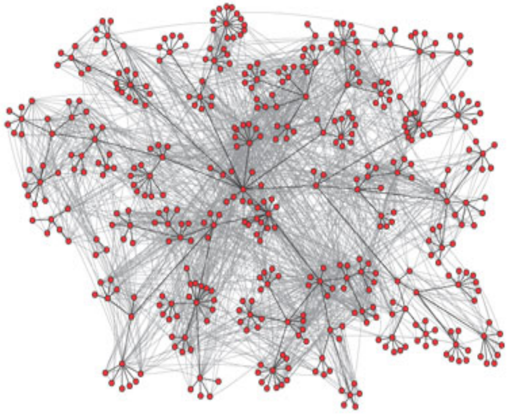
Network Properties

What do we observe that needs explaining?

- Small diameter (avg. shortest path length)
- High clustering coefficient
- Skewed degree distribution

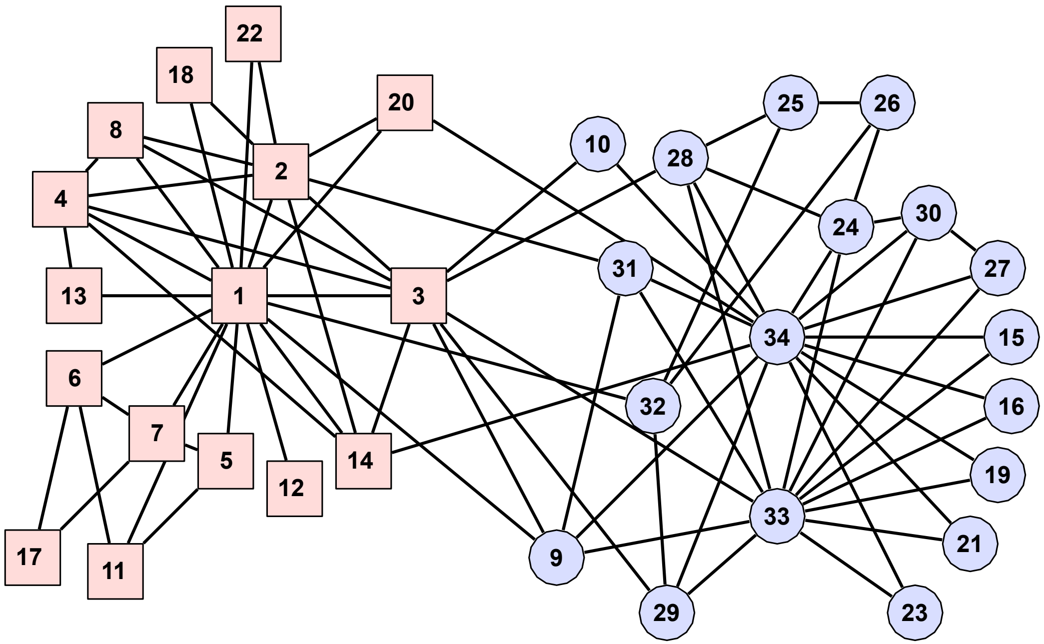
focus of this lecture

- Study power-laws
- explain the power-law degree distr. using the preferential attachment model

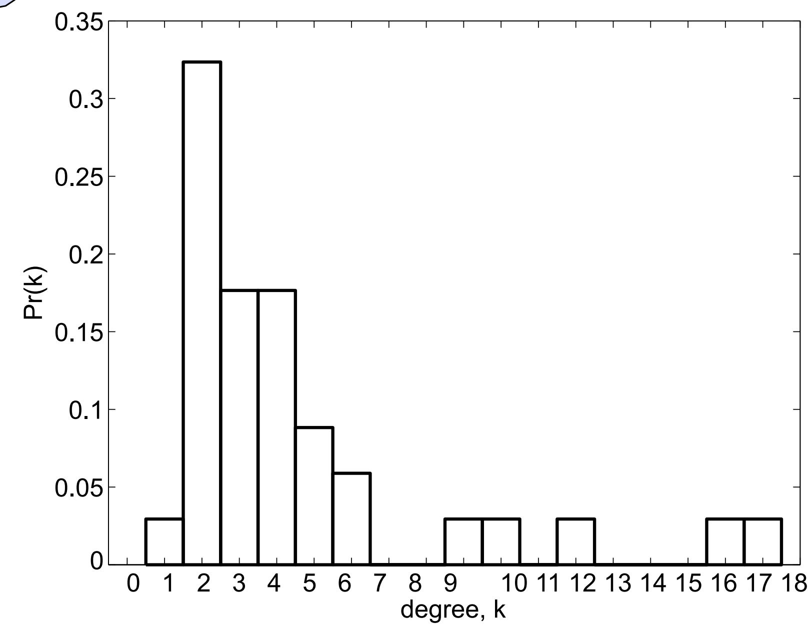


Power laws and Scale-free networks

Degree Distribution (1/3)

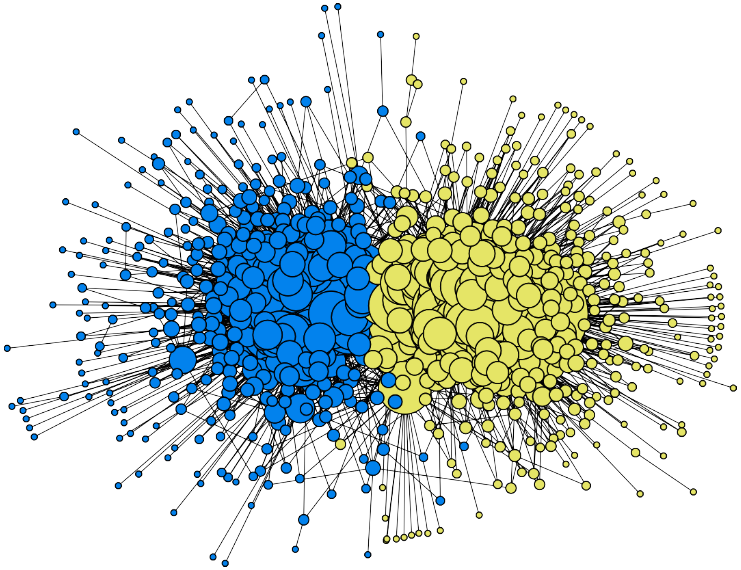


Zachary karate club

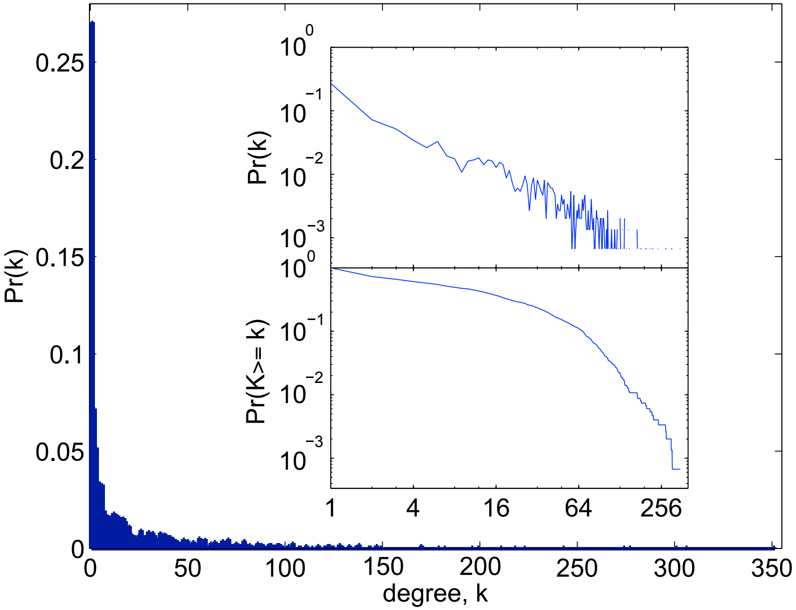


Source: A. Clauset, CU Boulder

Degree Distribution (2/3)



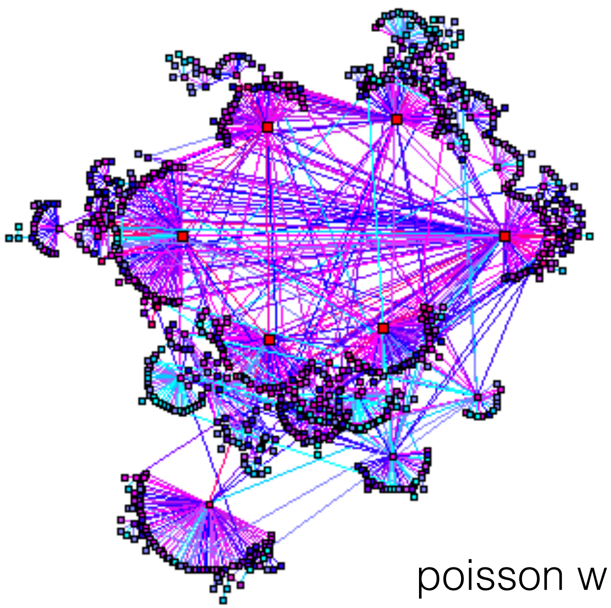
political blogs



Source: A. Clauset, CU Boulder

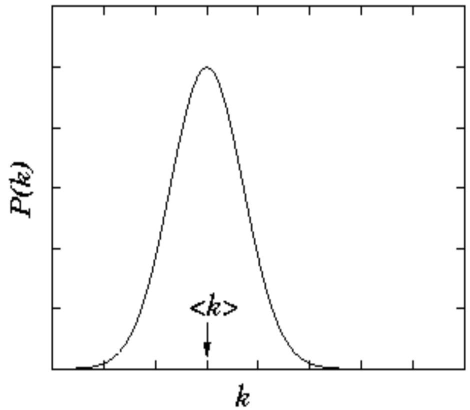
Degree Distribution (3/3)

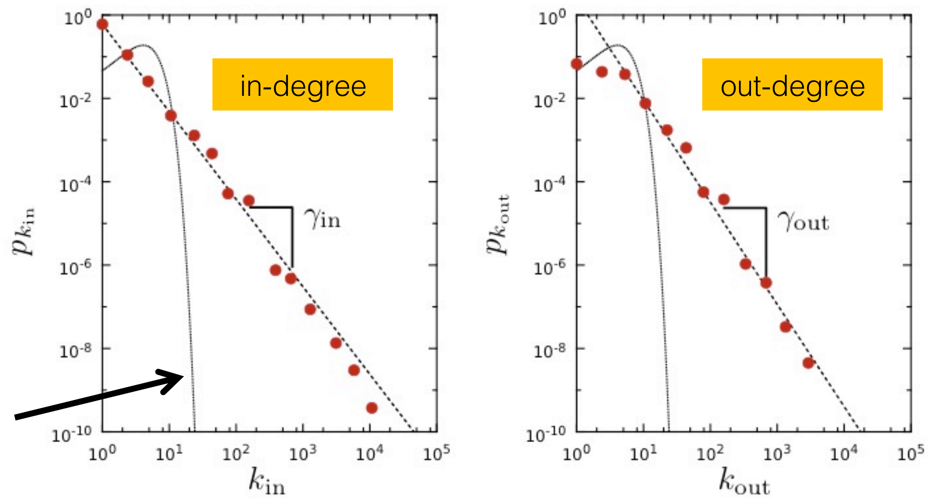
Nodes: WWW webpages
Edges: URL links



poisson with:
$\langle k_{in} \rangle = \langle k_{out} \rangle = 4.6$

Expected based on $G_{n,p}$



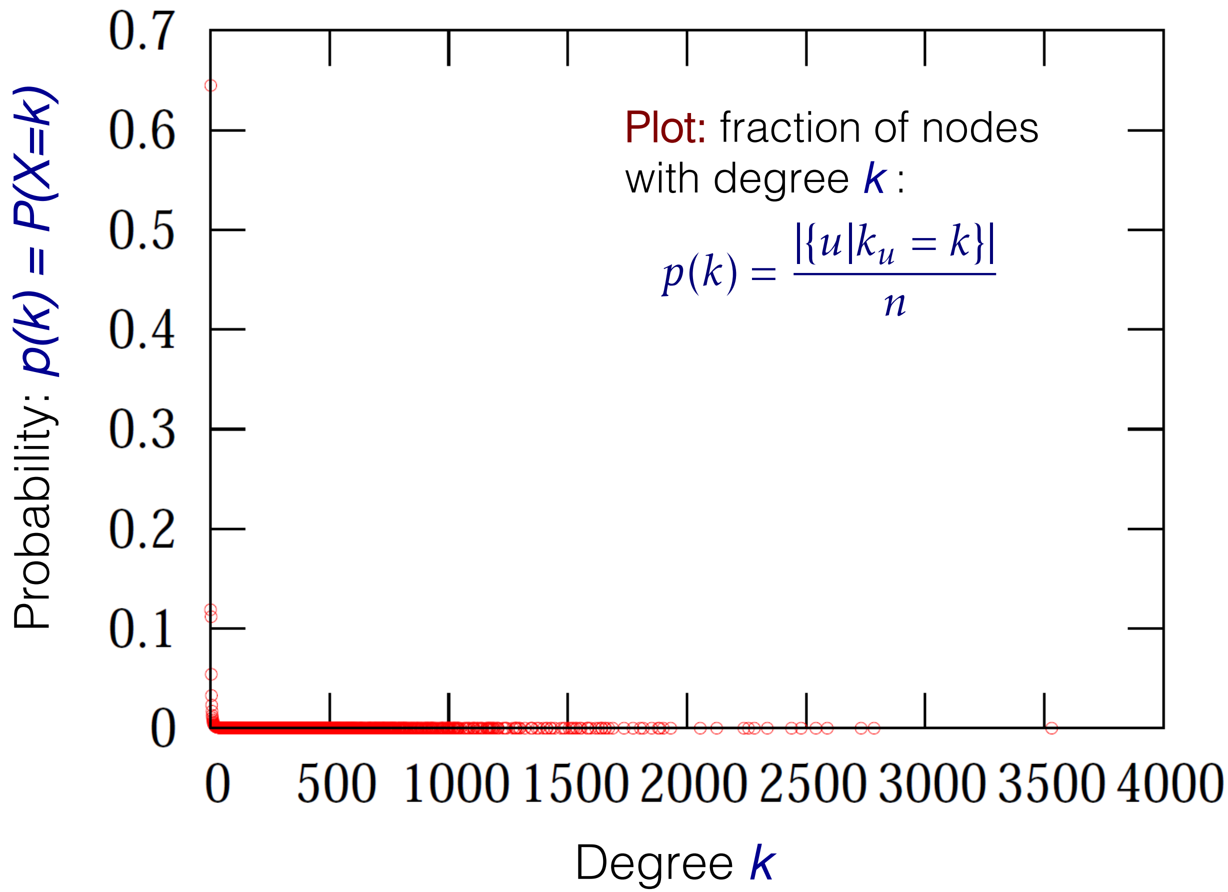


What we have observed from the data

R. Albert, H. Jeong, A-L Barabasi, Nature, 401 130 (1999)

Node Degrees in Networks (1/2)

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

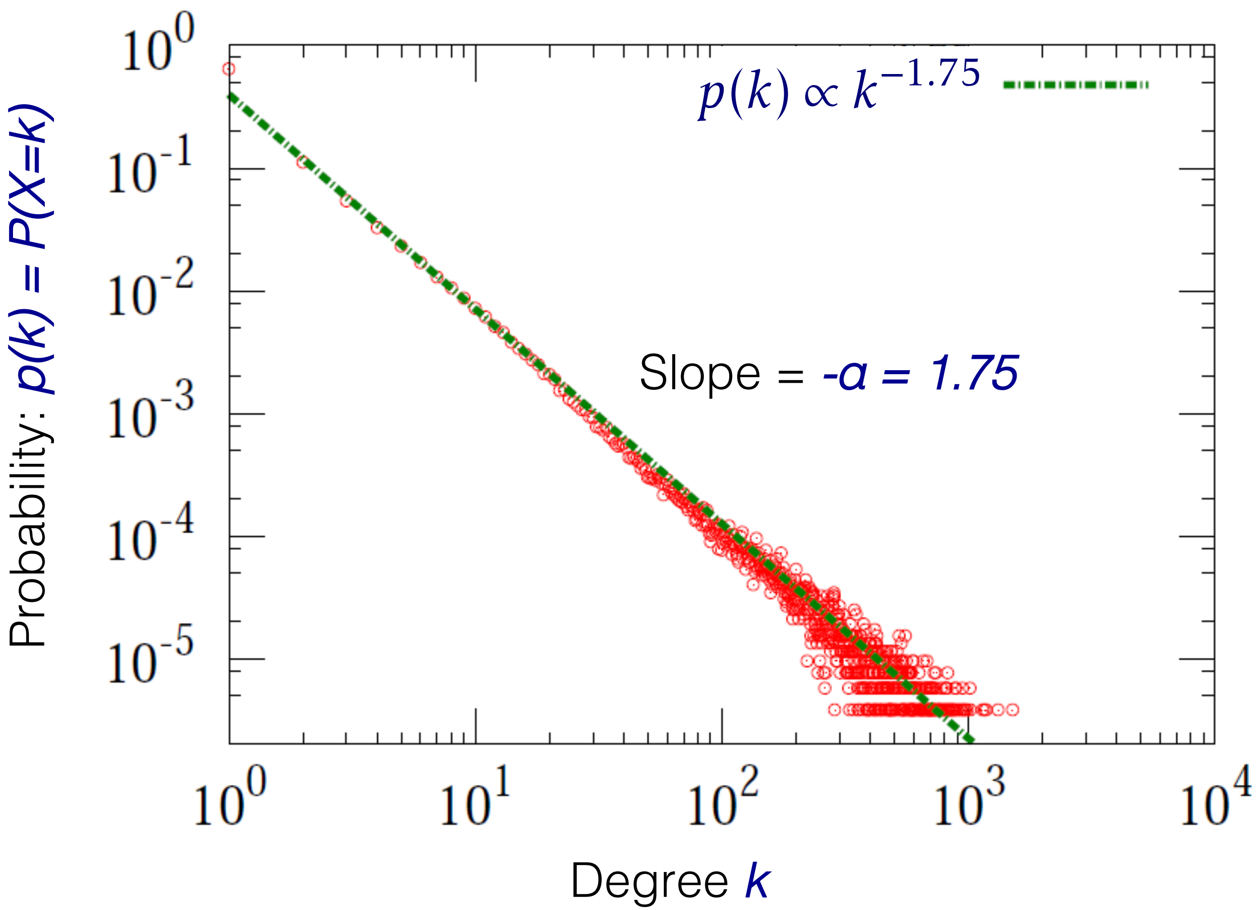


Flickr social network
n = 584,207
m=3,555,115

[Leskovec et al. KDD '08]

Node Degrees in Networks (2/2)

Plot the same in log-log scale



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

[Leskovec et al. KDD '08]

Power-law Degree Distribution

- Let $p(k)$ = fraction of nodes with degree k

$$p(k) \propto c\, k^{-\alpha}$$

with $\alpha > 1$ and c a constant

- How to recognize a power-law distribution?

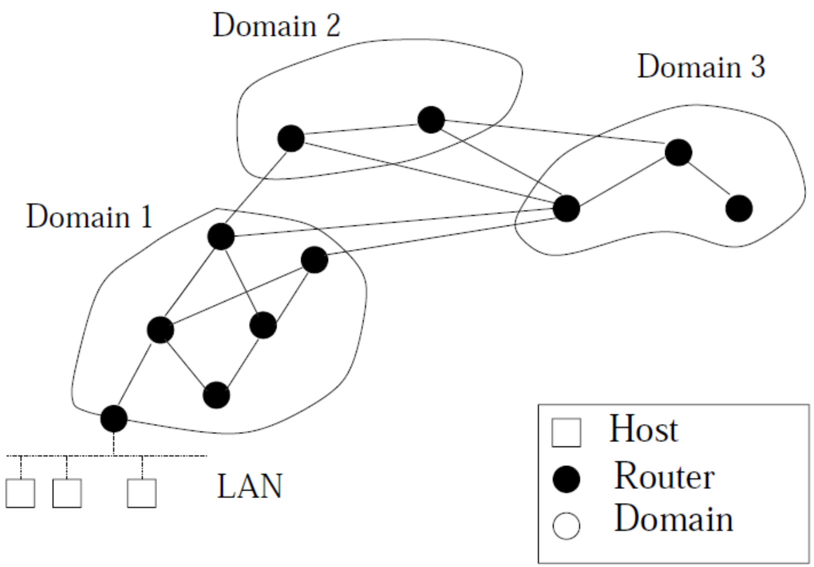
$$\log p(k) = \log c - \alpha \log k$$

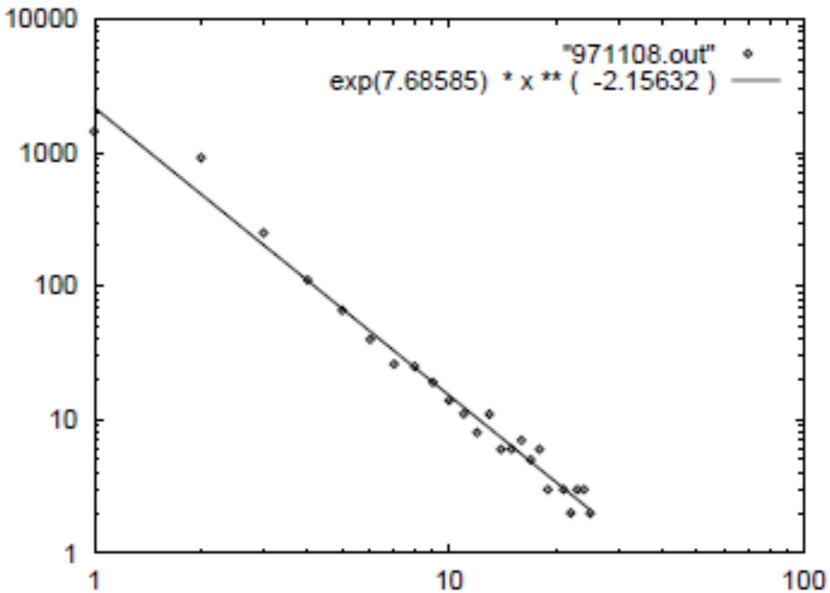
Plotting $\log p(k)$ versus $\log k$ gives a straight line with slope $-\alpha$

Node Degrees: Faloutsos[3]

Internet Autonomous Systems

[Faloutsos, Faloutsos and Faloutsos, 1999]

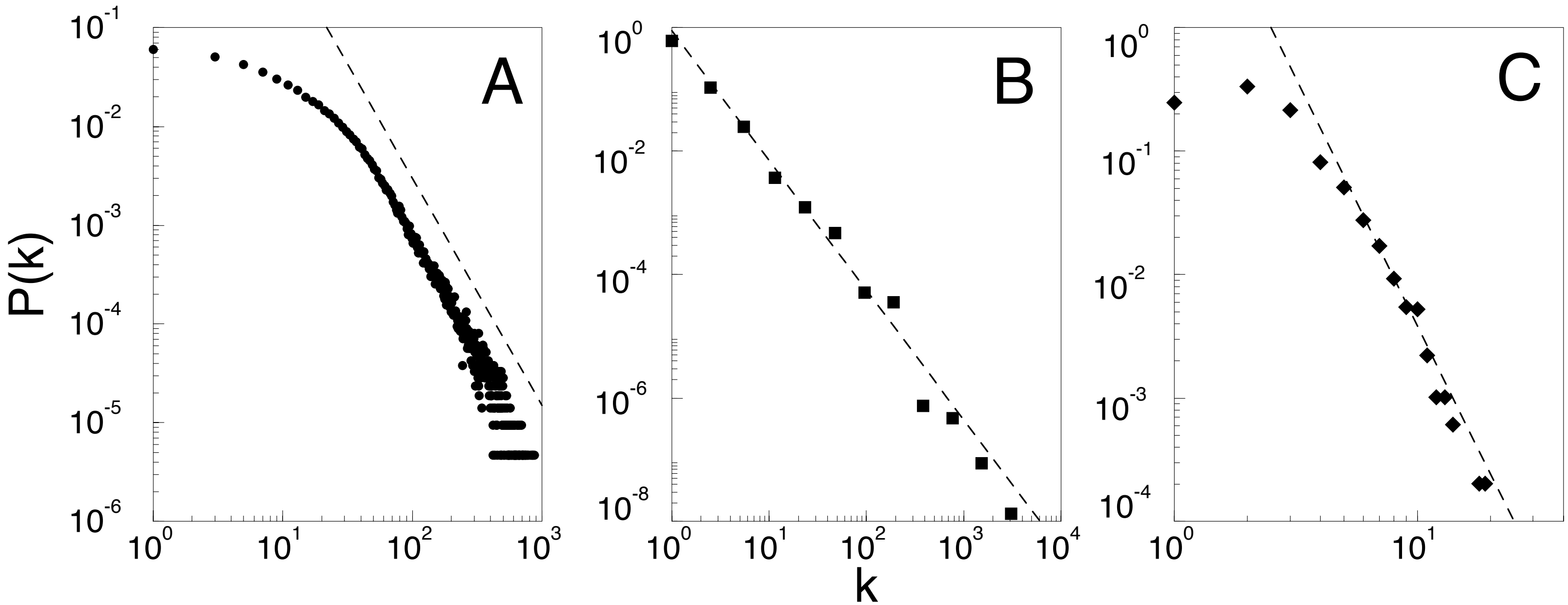




Internet domain topology

Node Degrees: Barabási and Albert

- Other Networks [Barabasi and Albert, 1999]

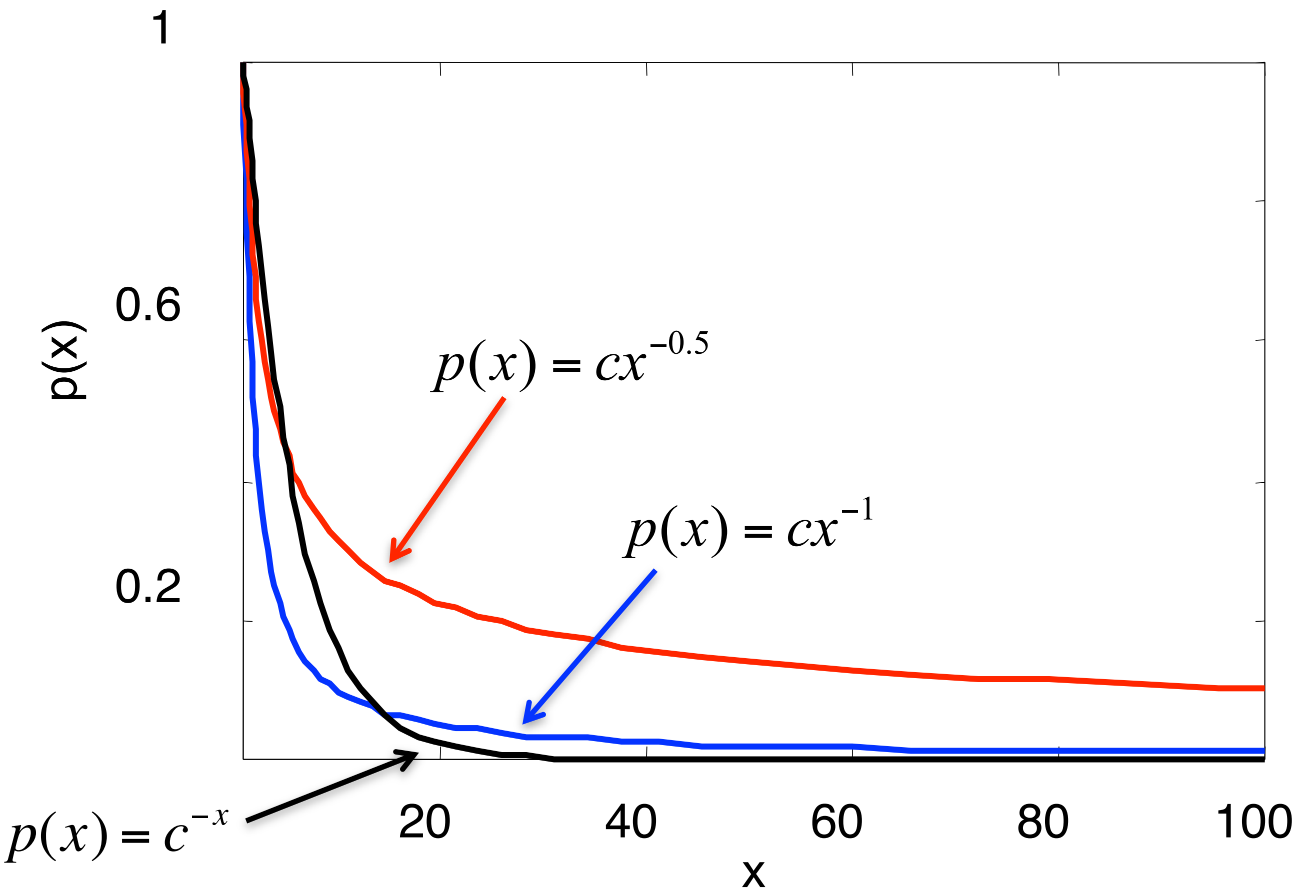


Actor collaborations

Web graph

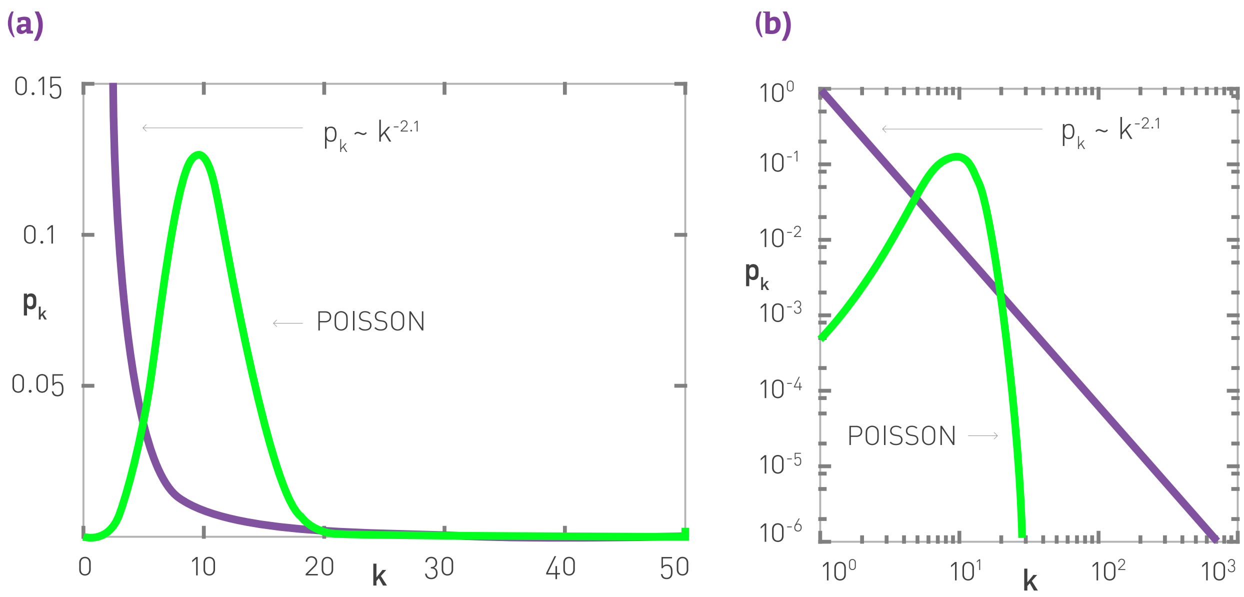
Power-grid

Exponential vs. Power-Law Distribution



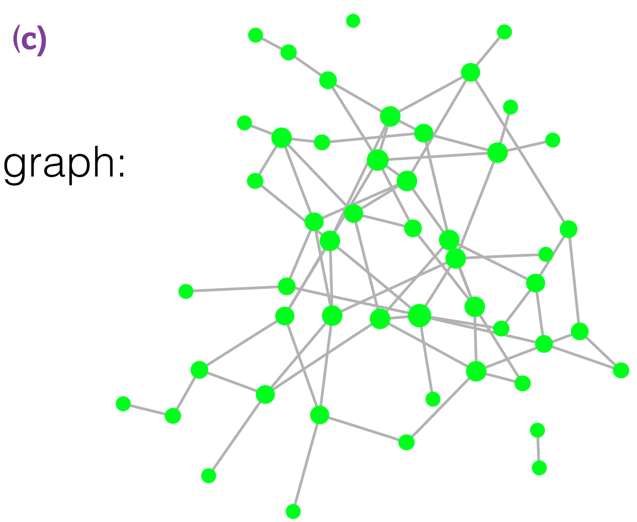
Above a certain value, the power law is always higher than the exponential!

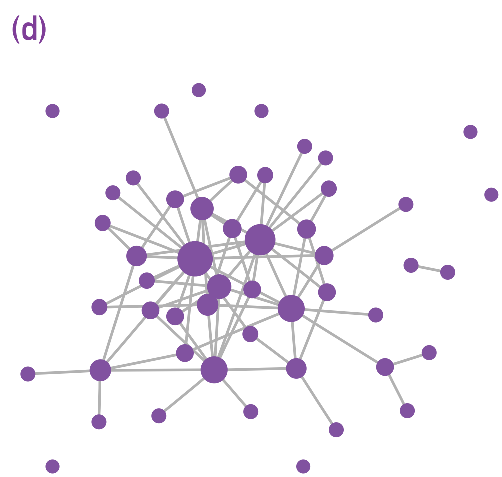
Power-Law vs. Exponential Degree Distributions



$\langle k \rangle = 11$

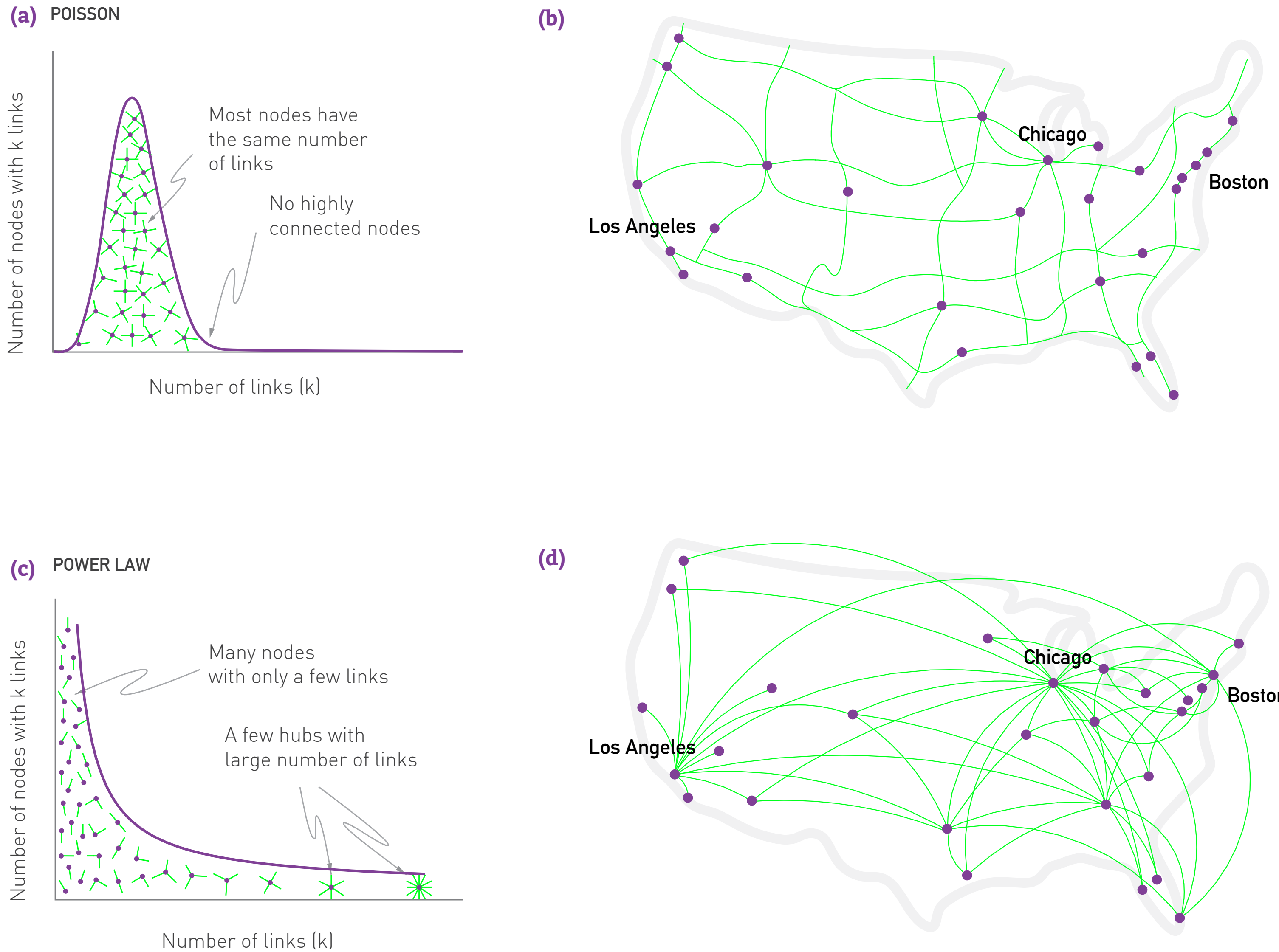
Random graph:
$\langle k \rangle = 3$
$n = 50$





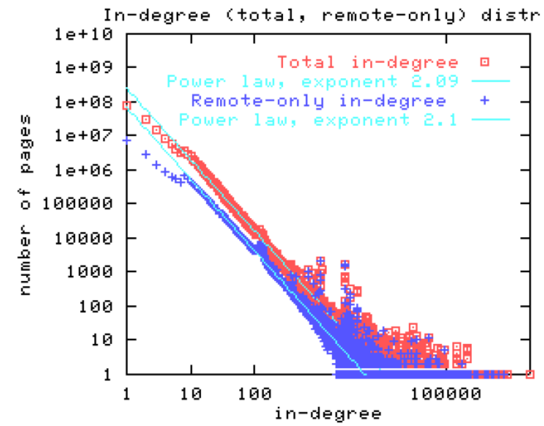
Power-law degree distr. graph:
$\langle k \rangle = 3$
$\alpha = 2.1$

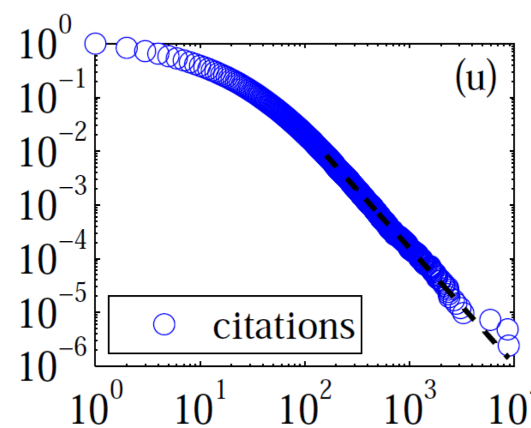
Power-Law vs. Exponential Degree Distributions



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]

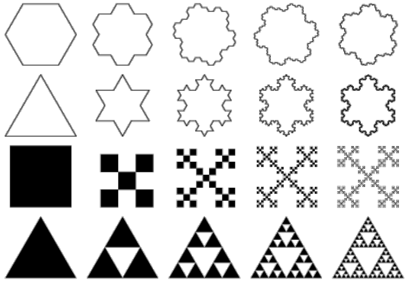




Scale-Free Networks

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

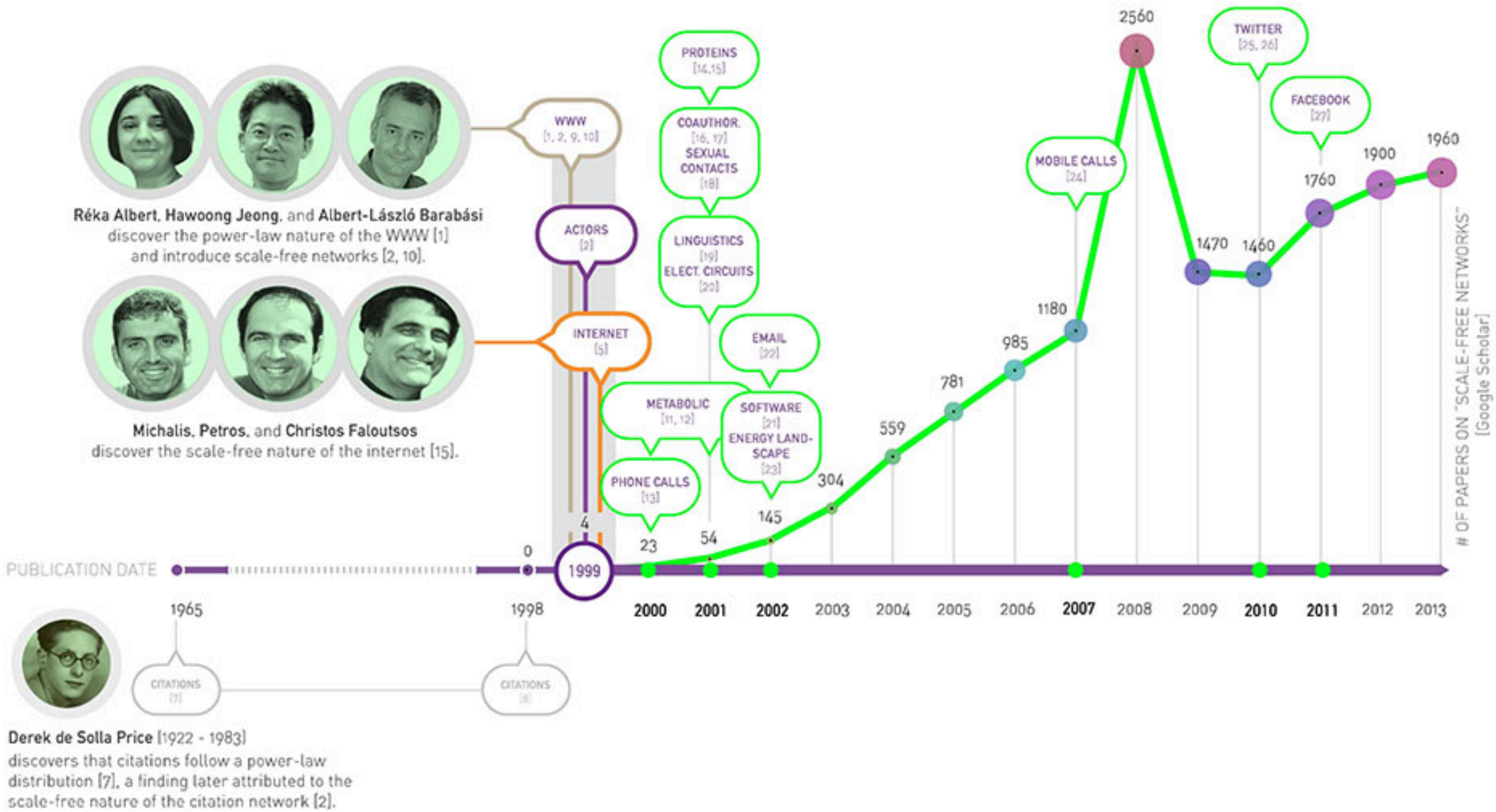
- Name motivated by the scale-invariance property of power-laws
  - Scale invariance: there is no characteristic scale
  - Laws do not change if scales of length, energy, or other variables, are multiplied by a common factor



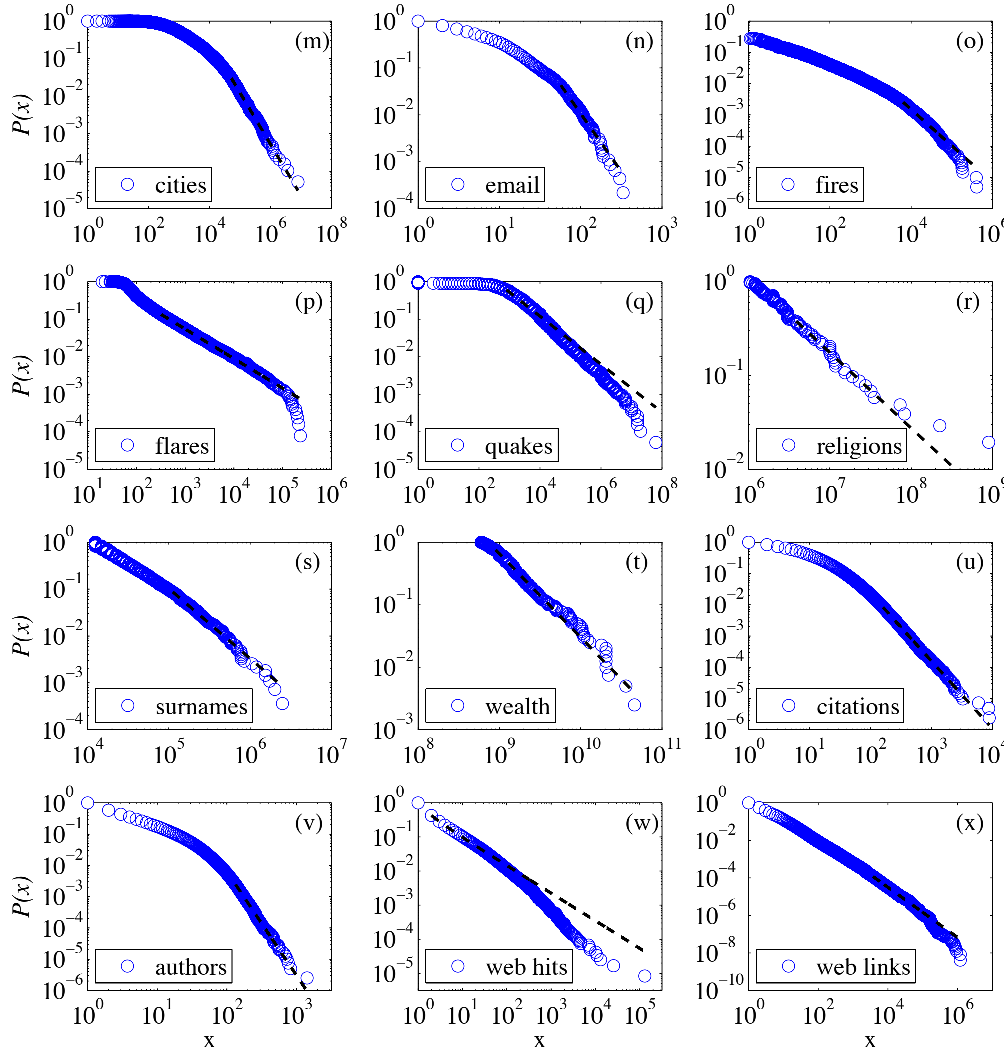
- Scale-free function $f(x)$ satisfies $f(\lambda x) = \lambda^{\Delta} f(x)$ for some scaling factor λ choice of exponent Δ
  - Power-law function: $f(x) = x^{-\alpha}$
    - $f(\lambda x) = (\lambda x)^{-\alpha} = \lambda^{-\alpha} f(x)$ it is scale-free
  - Exponential functions: $f(x)=c^x$
    - $f(\lambda x) = c^{\lambda x} = (c^x)^{\lambda} = f^{\lambda}(x) \neq \lambda^{\Delta} f(x)$ it is not scale-free

No characteristic scale for the degrees

Timeline of Scale-Free Networks



Power-Laws are Everywhere



[Clauset et al. 2009]

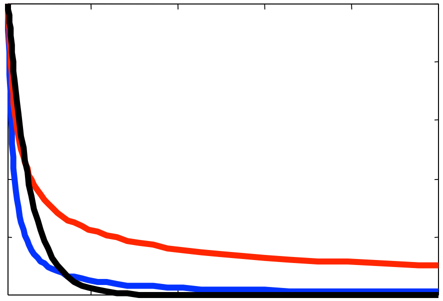
Many other quantities (beyond networks) follow heavy-tailed distributions

Not Everyone Likes Power-Laws ☺

CMU graduate students at the G20 meeting in Pittsburgh in Sept. 2009

Theoretical properties 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 $P(X > x) = 1 - P(X \leq x) = e^{-\lambda x}$

They are not heavy-tailed

Normalization

- Q: What is the normalizing constant *c* in $p(k) = ck^{-\alpha}$?

- *p(k) = p(X=k)* is a distribution, thus $\int p(k)dk = 1$

PDF

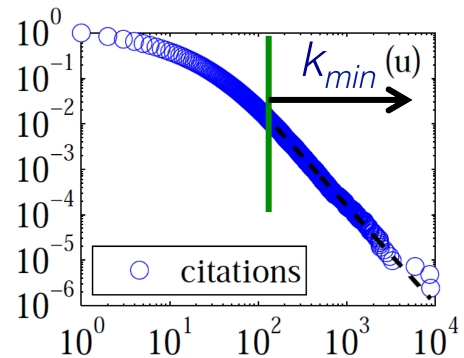
Continuous approximation

$$1 = \int_{k_{min}}^{\infty} p(k)dk = c\int_{k_{min}}^{\infty} k^{-\alpha}dk$$

$$= -\frac{c}{\alpha-1}\left[k^{-\alpha+1}\right]_{k_{min}}^{\infty} = -\frac{c}{\alpha-1}\left[\infty^{1-\alpha} - k_{min}^{1-\alpha}\right]$$

(Need: $\alpha > 1$)

$$\Rightarrow c = (\alpha-1)k_{min}^{\alpha-1}$$



Often, a power-law is only valid for the tail $k > k_{min}$

p(k) diverges as $k \to 0$
so k_{min} is the minimum value of the power-law distribution $k \in [k_{min}, \infty]$

Normalized power-law degree distribution:

$$p(k) = \frac{\alpha-1}{k_{min}}\left(\frac{k}{k_{min}}\right)^{-\alpha}$$

More on k_{min} soon!

Expected Value

- Q: What is the expected value of a power-law random variable K ?

$$\mathbb{E}[K] = \int_{k_{min}}^{\infty} k\ p(k)dk = c\int_{k_{min}}^{\infty} k^{-\alpha+1}dx$$

$$= \frac{c}{2-\alpha}\left[k^{2-\alpha}\right]_{k_{min}}^{\infty} = \frac{(\alpha-1)k_{min}^{\alpha-1}}{-(\alpha-2)}\left[\infty^{2-\alpha} - k_{min}^{2-\alpha}\right]$$

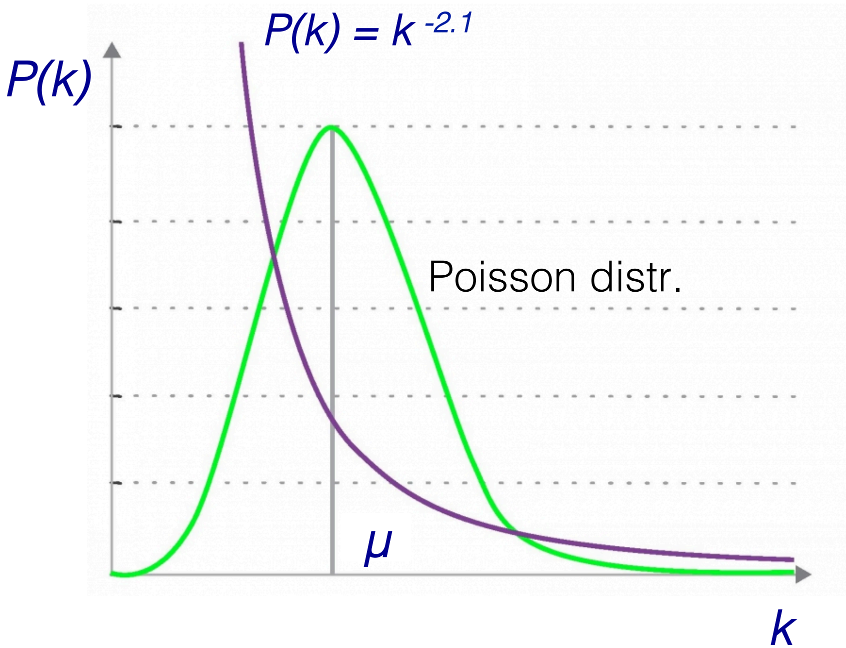
(Need: $\alpha > 2$)

$$= \frac{\alpha-1}{\alpha-2}k_{min}$$

In real networks
$2 < \alpha < 3$ so:
$E[K] < \infty$
$Var[K] = \infty$

Average is meaningless,
as the variance is too high

Scale-free Property (Again)



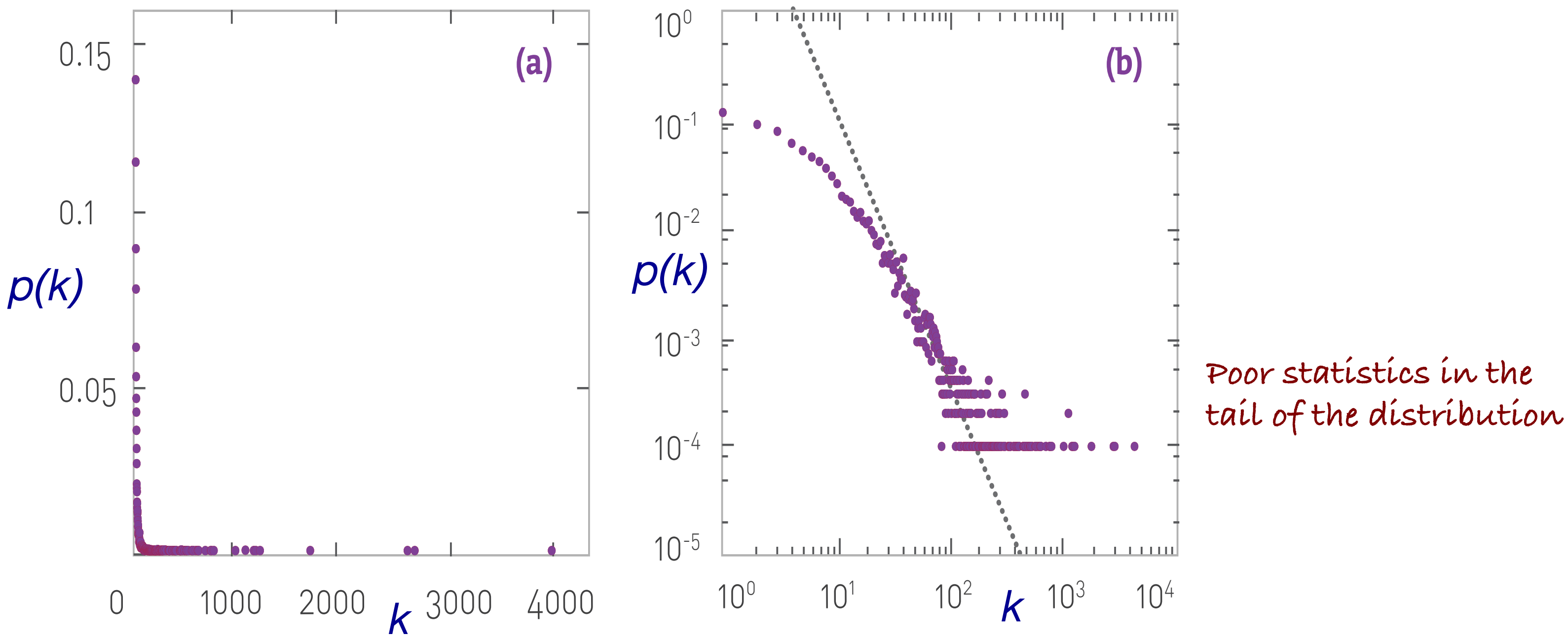
A measure of scale of a random variable is its standard deviation σ

- Large random graph $G_{n,p}$
  - Randomly chosen node has degree $k = \mu \pm \sqrt{\mu}$ The scale is μ
- Scale-free network
  - Randomly chosen node has degree $k = \mu \pm \infty$ There is no scale

Visualizing and fitting power-laws

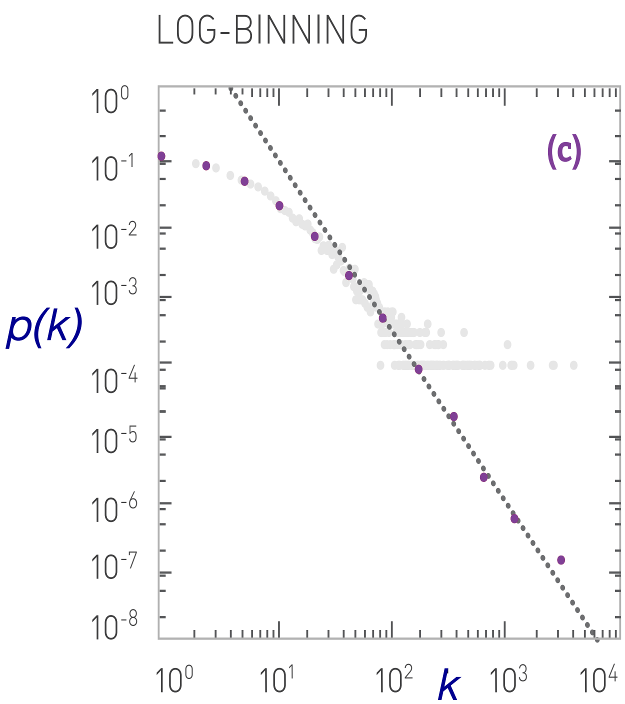
Visualizing Power-Law Distributions

- A simple histogram may not be enough for visualizing *p(k)*
  - Use log-log scale (figure (b))



- Large statistics fluctuations ('noisy signal') in the tail for large *k*
  - Increase the size (width) of the bins in the histogram
  - Larger bins contain more samples -> less noise (but also less details)

Logarithmic Binning



- We would like to use bins of different sizes in different parts of the histogram
  - Use increasingly larger bins as we go further out in the tail
  - Make each bin little larger that the previous one

- Apply logarithmic binning: each bin is made wider than its predecessor by a constant factor a
  - First bin: $1 \leq k < 2$ and $a = 2$
  - Second bin: $2 \leq k < 4$
  - n-th bin: $a^{n-1} \leq k < a^n$

Then, normalize each bin by its width

Complementary Cumulative Distribution Function

Another way to visualize a power-law distribution

- The complementary cumulative distribution function (CCDF) is

$$F(k) = P(K \geq k)$$

  - Function $F(k)$ is the fraction of vertices with degree at least k

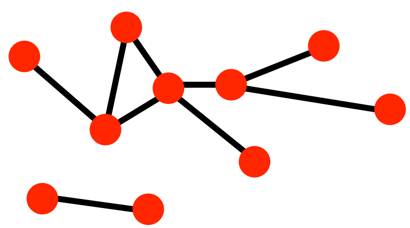
- If the PDF is power-law, the CCDF is also a power-law

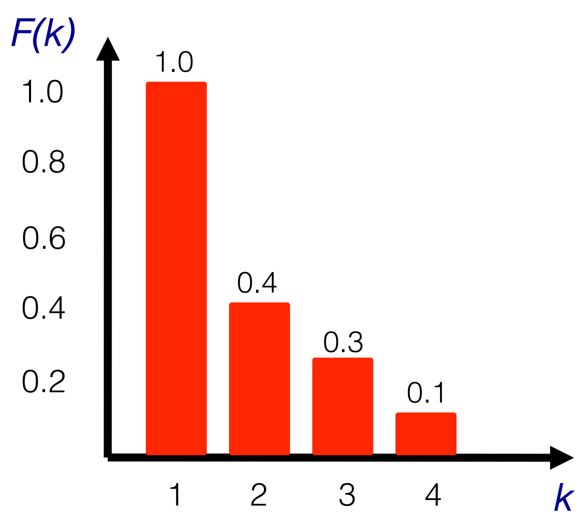
$$P(K \geq k) = \int_k^\infty \underbrace{\frac{\alpha - 1}{k_{min}} \left(\frac{k}{k_{min}}\right)}_{\text{PDF}} dk = \left(\frac{k}{k_{min}}\right)^{-(\alpha-1)}$$

If the PDF has exponent α, then the CCDF $F(k)$ has exponent α-1

Computing the CCDF

- Step 1: List the degrees k of the nodes in descending order
- Step 2: Assign ranks r_i (from 1 to $n=|V|$) to vertices in that order
- Step 3: The CCDF is the plot of r_i / n versus the degree k

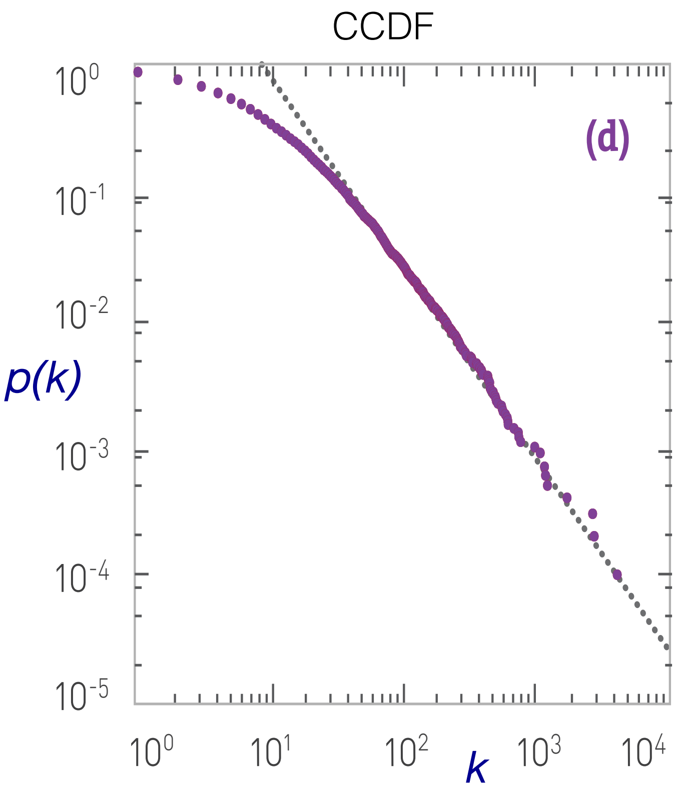




Degree k	Rank r_i	$F(k) = r_i / n$
4	1	0.1
3	2	0.2
3	3	0.3
2	4	0.4
1	5	0.5
1	6	0.6
1	7	0.7
1	8	0.8
1	9	0.9
1	10	1.0

If degrees are repeated, keep the largest value

Visualizing Power-Laws with CCDF

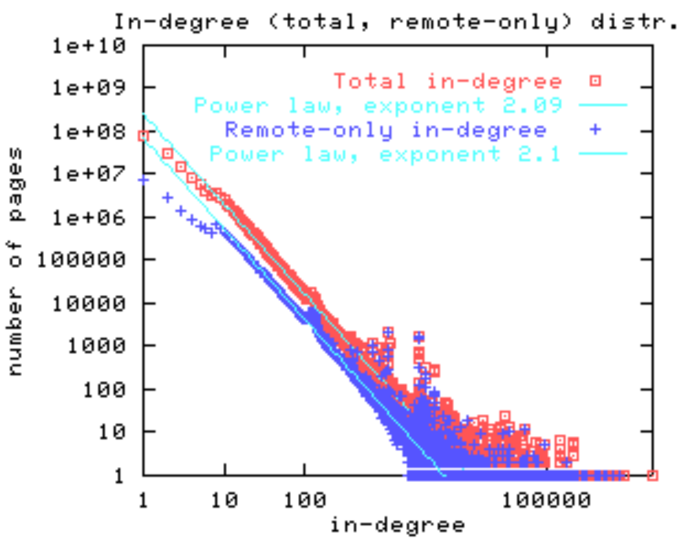


- Plot CCDF in log-log scale
- Pros
  - Reduces noise using cumulative frequencies (instead of raw frequencies)
  - No binning is needed
- Cons
  - Points in CCDF are correlated
  - Not a very good approach to use CCDF to find the exponent α of the power-law

Next topic

Fitting Power-Law Distributions (1/3)

- How to estimate the exponent α from the data



A power-law implies that: $$\log p(k) = C - \alpha \log k$$

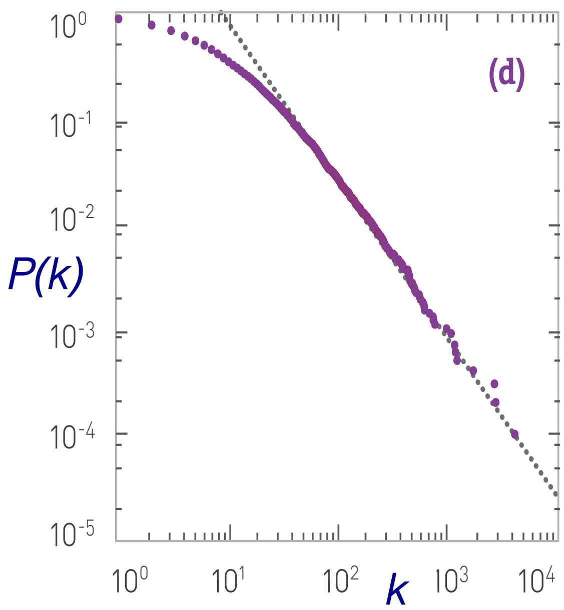
Linear least-squares regression

$$\arg\min_{\alpha} \sum_i \left(\log p(k_i) - C + \alpha \log k_i \right)^2$$

- Very simple, but not advisable approach
  - Noisy high-degree data
  - Typically, the power-law is only valid in the tail -> we need to pick k_{min}

Fitting Power-Law Distributions (2/3)

- Overcome noise using the CCDF
  – Cumulative frequencies smoothen the noise



- Recall that, the CCDF follows a power-law with exponent *α-1*
  – Use linear regression to estimate *α'*

- Good method, but has some disadvantages
  – Successive points in the CCDF are correlated
  – The least squares regression assumes independence between the data points

Fitting Power-Law Distributions (3/3)

Maximum-likelihood estimator

- Recall that the data PDF is $p(k) = \frac{\alpha - 1}{k_{min}} \left(\frac{k}{k_{min}} \right)^{-\alpha}$ for $k \geq k_{min}$
- Given a set of observations $\{k_i\}$ find α that is most likely to have generated our data

$$p(k|\alpha) = \prod_{i=1}^{n} \frac{\alpha - 1}{k_{min}} \left(\frac{k_i}{k_{min}} \right)^{-\alpha}$$

log-likelihood

$$\ln \mathcal{L}(\alpha) = \ln p(k|\alpha) = \ln \prod_{i=1}^{n} \frac{\alpha - 1}{k_{min}} \left(\frac{k_i}{k_{min}} \right)^{-\alpha}$$

$$= \sum_{i=1}^{n} \left(\ln(\alpha - 1) - \ln k_{min} - \alpha \ln \left(\frac{k_i}{k_{min}} \right) \right)$$

$$= n \ln(\alpha - 1) - n \ln k_{min} - \alpha \sum_{i=1}^{n} \ln \left(\frac{k_i}{k_{min}} \right)$$

Maximum Likelihood Estimator (MLE)

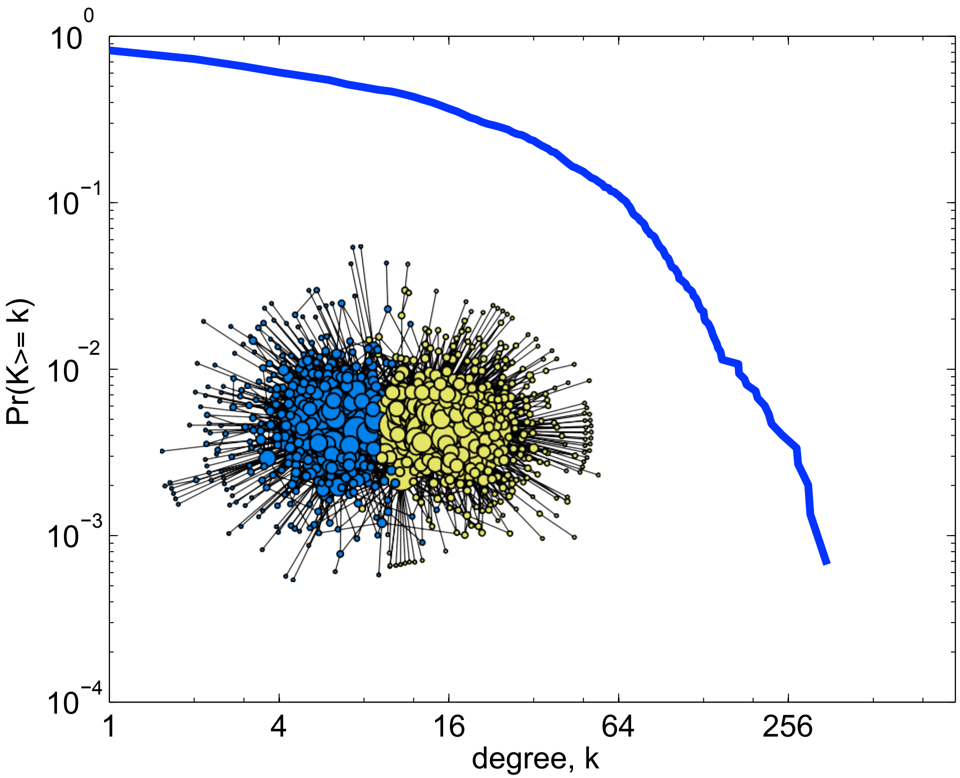
Set:

$$\frac{\partial \mathcal{L}}{\partial \alpha} = 0 \Rightarrow \frac{n}{\alpha - 1} - \sum_{i=1}^{n} \ln \left(\frac{k_i}{k_{min}} \right) = 0$$

Thus, the MLE estimator will be:

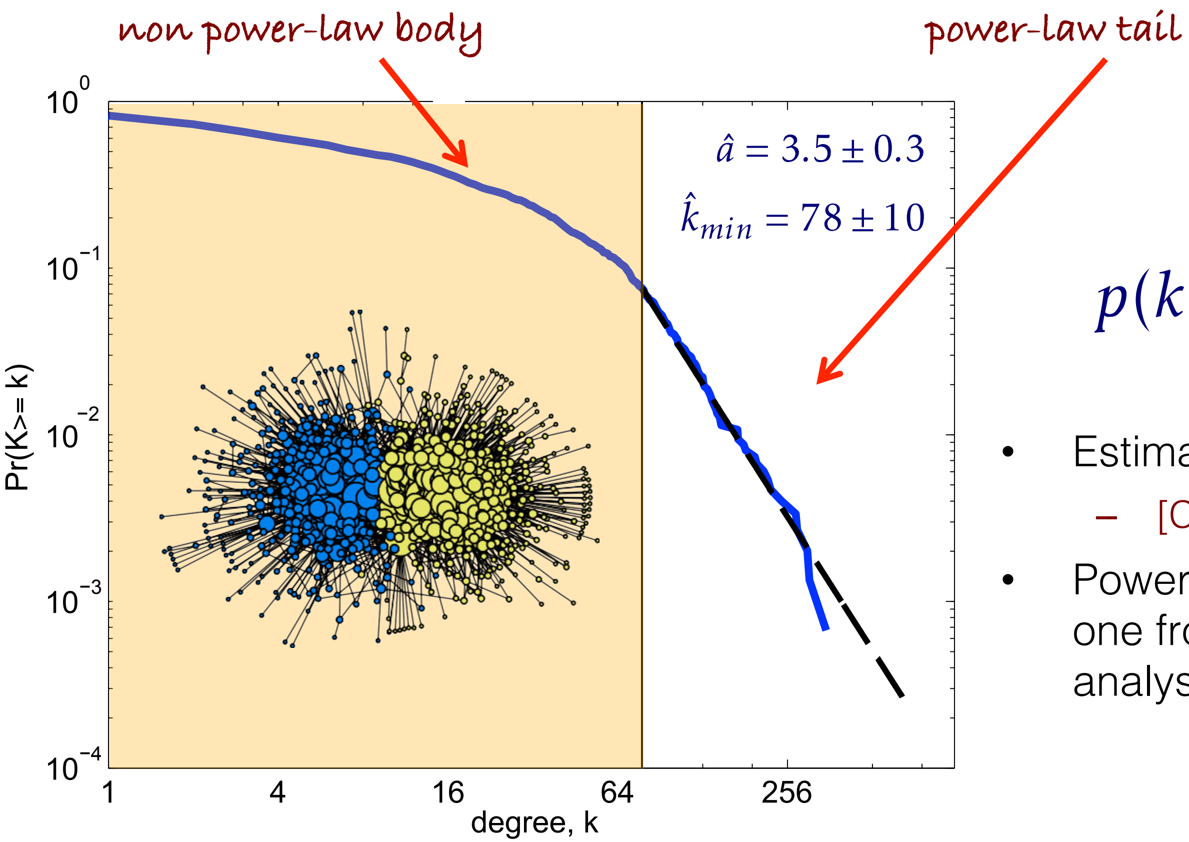
$$\hat{a} = 1 + n \left(\sum_{i=1}^{n} \ln \left(\frac{k_i}{k_{min}} \right) \right)^{-1}$$

Summary



Source: A. Clauset, CU Boulder

Summary



$$p(k) \propto c\, k^{-\alpha} \quad \text{for} \quad k \geq k_{min}$$

- Estimate k_{min} from data
  - [Clauset et al. 2009]
- Power-laws are cool! But knowing one from garbage needs good analysis

Toolkit for fitting power-law distributions in empirical data:

http://tuvalu.santafe.edu/~aaronc/powerlaws/

Source: A. Clauset, CU Boulder

The Preferential Attachment model

Growth and Preferential Attachment

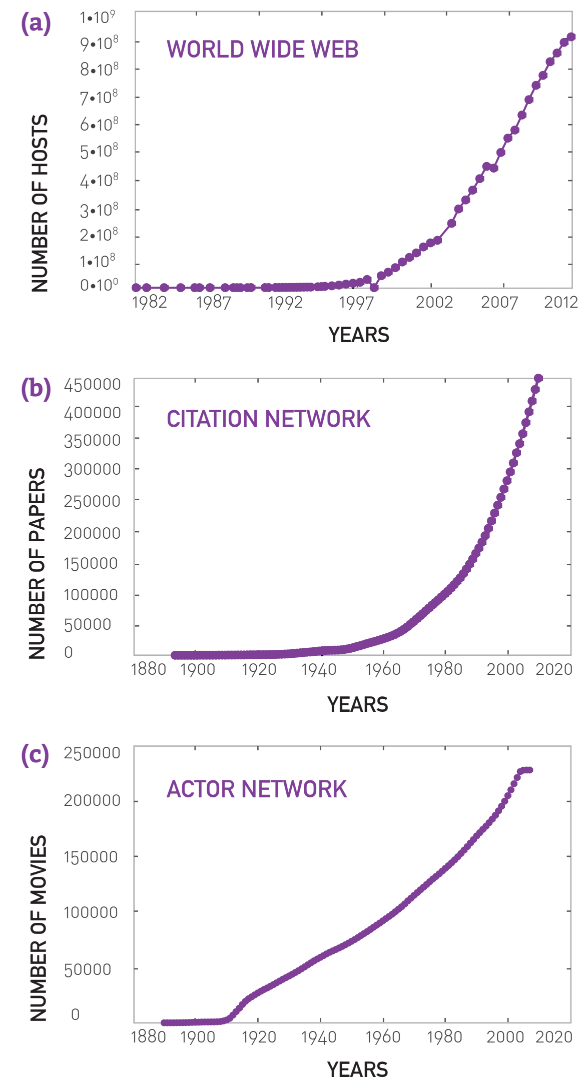
- Erdős–Rényi random graph model
  - The number of nodes *n* is fixed (static model)
  - Nodes randomly choose other nodes to interact

Growth

Networks expand through the addition of new nodes and edges

Preferential attachment

In real networks, nodes prefer to link to more connected nodes

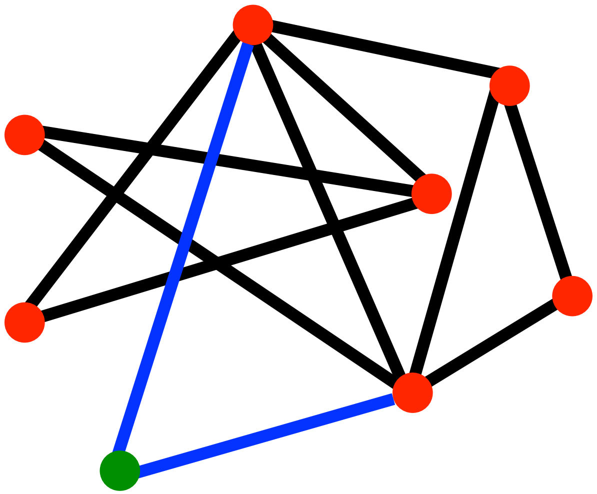


Model: Preferential attachment

- Preferential attachment
  - [Start with a small number m_0 of nodes]
  - Nodes arrive in order $1,2,\ldots,n$
  - At step j, let k_i be the degree of node $i < j$
  - A new node j arrives and creates m links
  - Prob. of j linking to a previous node i is proportional to the degree d_i of node i

$$P(j \to i) = \frac{k_i}{\sum_v k_v}$$

All pre-existing nodes v



[de Solla Price '65, Albert and Barabasi '99, Mitzenmacher '03]

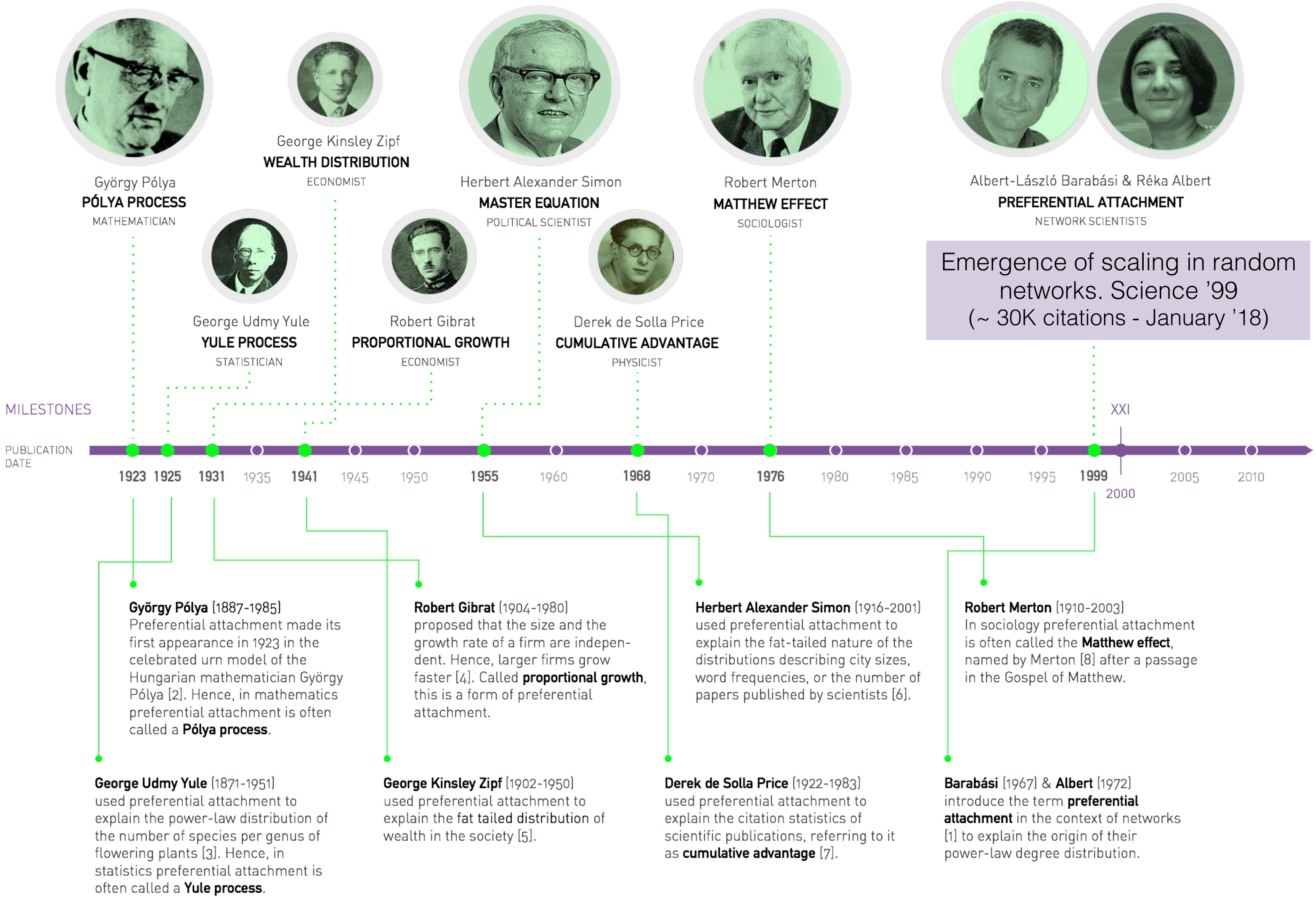
Rich-Get-Richer Effect

New nodes are more likely to link to nodes that already have high degree

- Herbert Simon (Economist, Sociologist 1916-2001)
  - Power-laws arise from the "**Rich-get-richer**" effect (cumulative advantage)

- Examples
  - Citations: New citations to a paper are proportional to the number it already has [de Solla Price '65]
    - If many people cite a paper, then it must be good, and therefore I should cite it too
  - Sociology: Matthew effect
    - Eminent scientists often get more credit than a comparatively unknown researcher, even if their work is similar
    - http://en.wikipedia.org/wiki/Matthew_effect

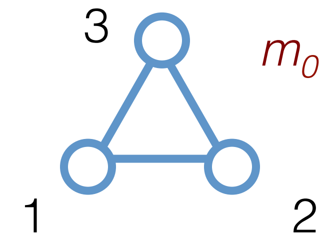
Preferential Attachment: A Brief History



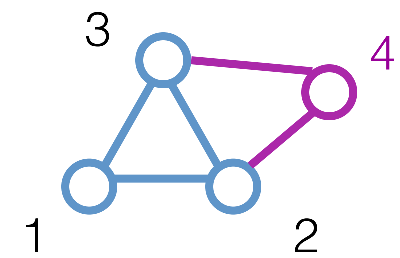
The Barabási-Albert Model

- To start, each node has an equal number of edges (= 2)
  - The probability of choosing any vertex is 1/3

1 1 2 2 3 3

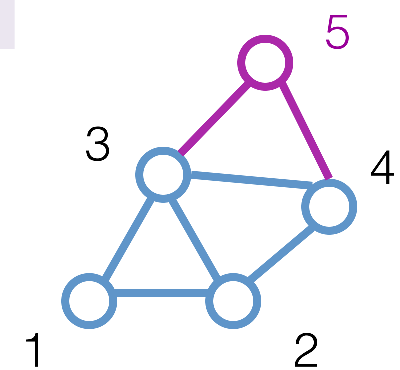


- We add a new node 4, and it will have $m=2$ edges
  - Draw 2 random elements from the array – suppose they are 2 and 3



1 1 2 2 2 3 3 3 4 4

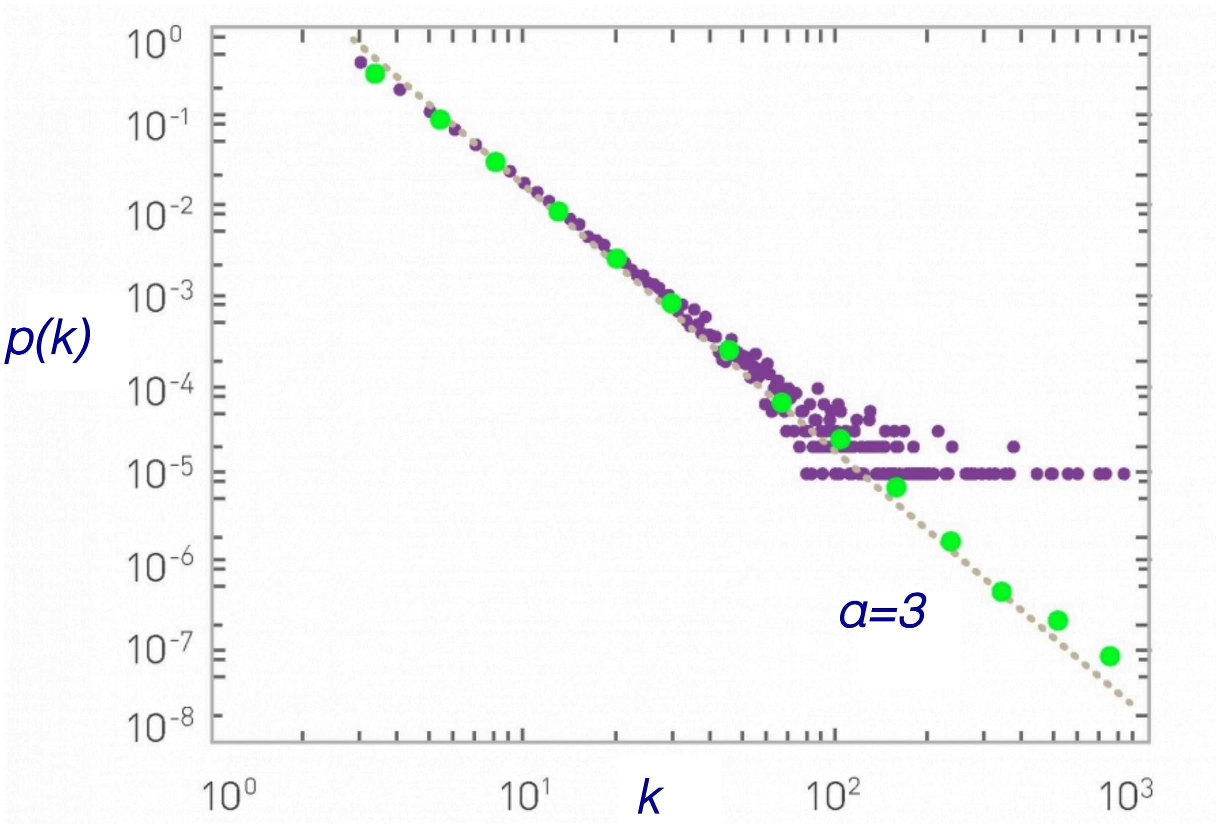
- Add a new node 5
  - Now the probabilities of selecting 1,2,3,or 4 are 1/5, 3/10, 3/10, 1/5



After t timesteps, the BA model generates networks with $n_t = t+m_o$ nodes and m_0+mt edges

1 1 2 2 2 3 3 3 4 4 4 5 5

Degree Distribution of BA Model



n = 100,000
m = 3

Power-law degree distribution: $p(k) \propto k^{-3}$

Evolution of the BA Model

Time for a short video:

https://www.youtube.com/watch?v=4GDqJVtPEGg

Some Aspects of the BA Model

- It does not specify the precise initial configuration of the first m_0 nodes
- It does not specify whether the m links assigned to a new node are added one by one, or simultaneously
  - If the links are truly independent, they could connect to the same node i, resulting in multi-links
- The power-law exponent of the degree distribution is fixed at $\alpha = 3$
- The BA model generates only undirected graphs
- It correlates age with degree which is not always the case
- Several extensions of the model
  - Linearized Chord Diagram (LCD)
  - Model with initial attractiveness
  - Copying mechanism

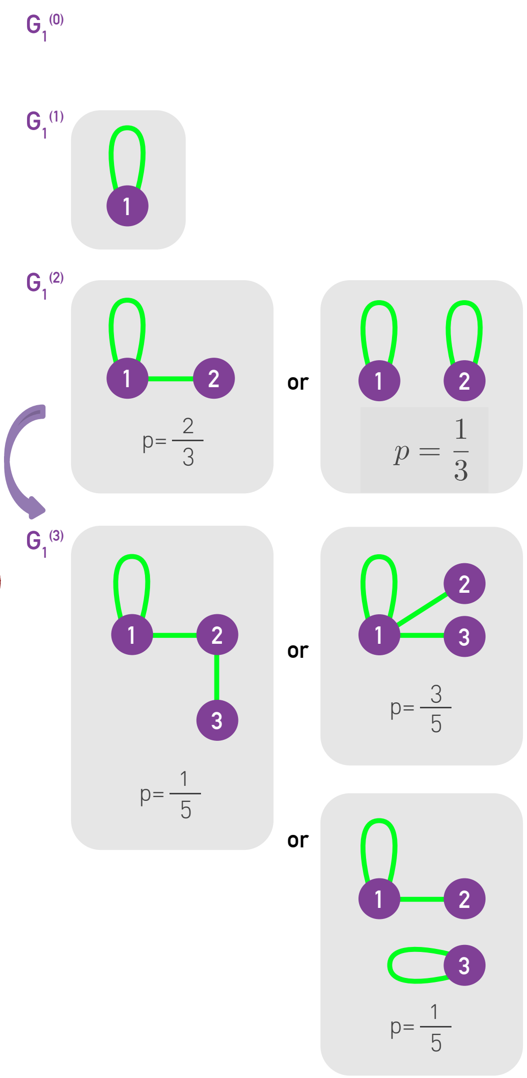
Linearized Chord Diagram (LCD)

For $m = 1$, we build a graph $G_1(t)$ as follows:

- Start with $G_1(1)$ consisting of a node with a self-loop
- For $t = 1,2,3,\ldots$ the current graph $G_1(t\text{-}1)$ grows to $G_1(t)$ as follows:
  - Add new node v_t and a single edge between v_t and v_i, where v_i is chosen with probability

$$p(v_i) = \begin{cases} \frac{k_i}{2t-1} & \text{if } 1 \leq i \leq t-1 \quad \textit{for the existing nodes} \\ \frac{1}{2t-1} & \text{if } i = t \quad \textit{self-loops} \end{cases}$$

- For $m>1$, run the above process m times for each t
  - Collapse all created vertices into a single one, retaining edges



[Bollobás et al. '01]

Properties of LCD Model

- The LCD model allows for loops and multi-edges
  - When $t \to \infty$ their number becomes negligible
- The LCD models produces graph with power-law degree distribution and *α = 3*
- The BA models produces connected graphs
  - Not true for the LCD model, but $G_1(t)$ will be connected w.h.p.
- Small-world behavior

$$\mathrm{diam}(G_1(t)) = \begin{cases} O(\log n_t) & \text{if } m = 1 \\ O(\frac{\log n_t}{\log \log n_t}) & \text{if } m > 1 \end{cases}$$

n_t is the number of nodes at time t

- Not good clustering coefficient

$$\mathbb{E}\Big(C(G_1(t))\Big) \approx \frac{m-1}{8} \frac{(\log n_t)^2}{n_t}$$

  - Marginally better than in random networks

Strong mathematical formulation of the BA model

Generate BA Graphs with $\alpha = [2, \infty)$

- For the BA model, we have that ***α = 3***
- Can we generate graphs with $\alpha = [2, \infty)$?
  - Modify the BA model by adding an extra parameter ***A≥0***

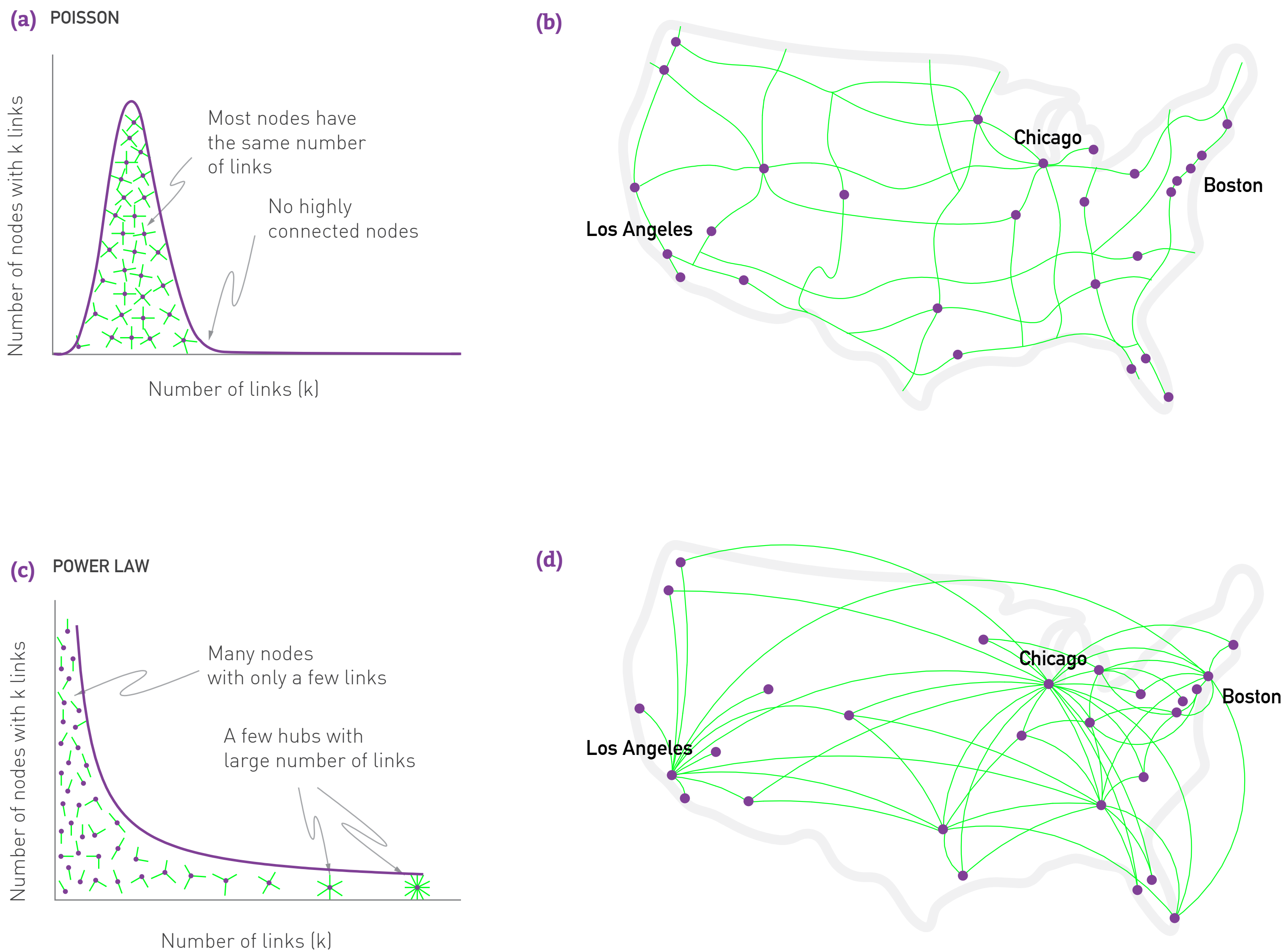
$$P(j \to i) = \frac{A + k_i}{\sum_v A + k_v}$$

  - Parameter ***A*** models the initial attractiveness of each node
  - If ***A=0***, then the model is the same as the BA model

The degree distribution follows a power-law with exponent: $\alpha = 2 + \frac{A}{m}$

[Dorogovtsev et al. '00]

Any Consequences of this Structural Property?

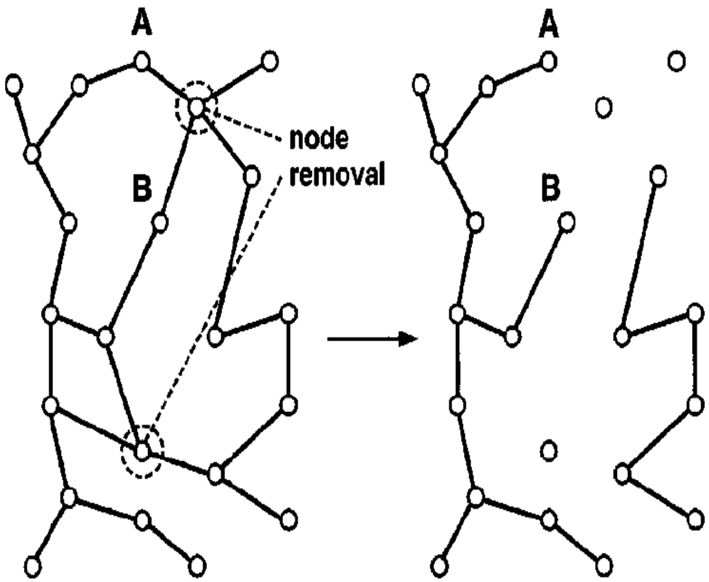


Consequences of power-law degree distribution

Error and attack tolerance of complex networks

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 can have an important effect in the robustness of the internet as well as in the domain of epidemiology

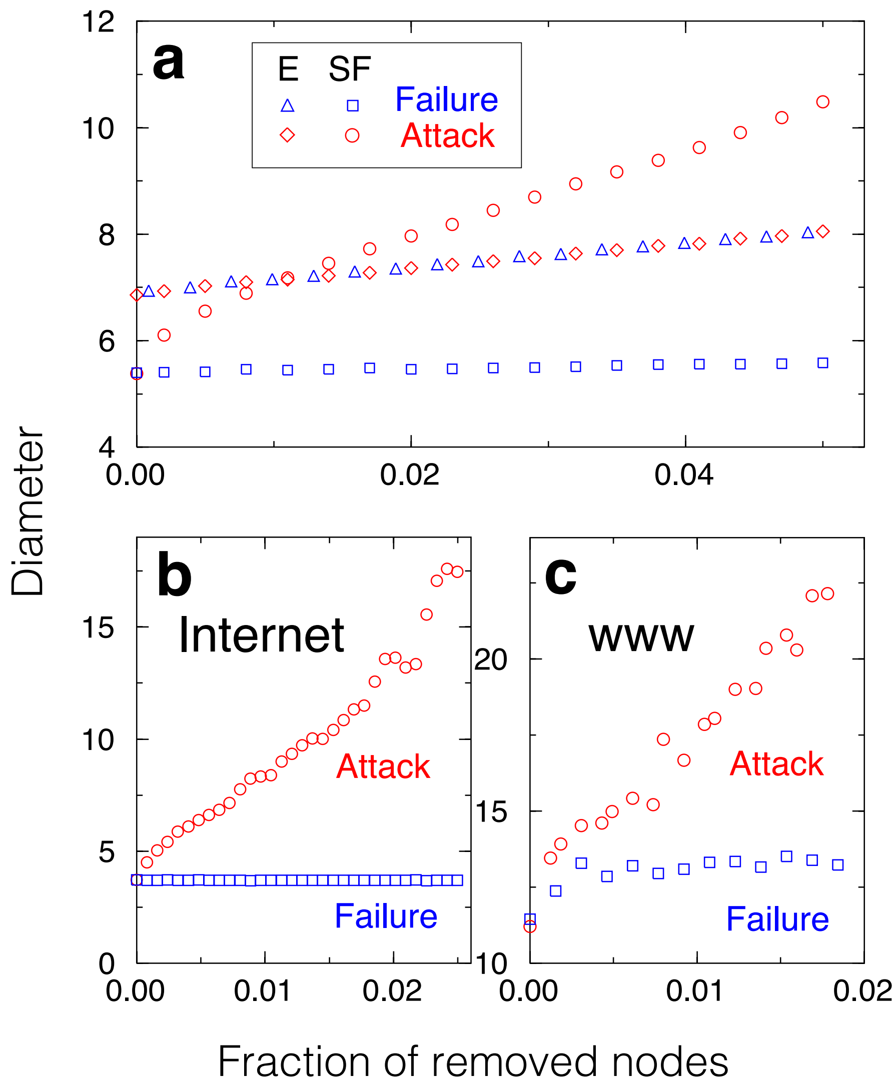
Achilles' Heel of Complex Networks

https://en.wikipedia.org/wiki/Achilles%27_heel

- Cover of the 27 July 2000 issue of Nature
- Highlight of the paper "Attack and error tolerance of complex networks" by R. Albert, H. Jeong and A.-L. Barabási
    - Network robustness

How to quantify network robustness?

Error Tolerance



E: Random network (Exponential degree distr.)
SF: Scale-free network (Power-law degree distr.)

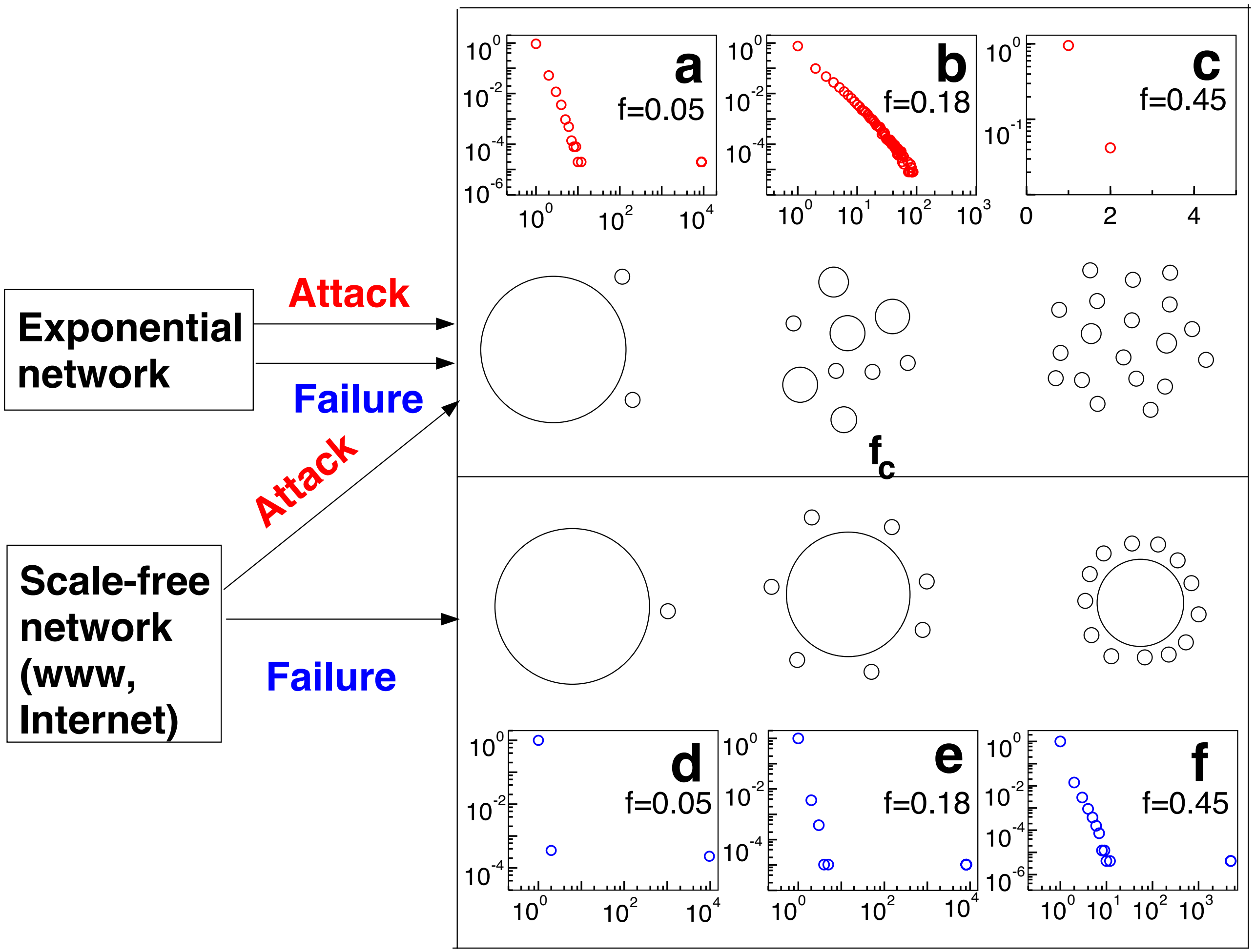
How does the diameter change if we remove a fraction of nodes

- Randomly (Failure)
- Targeted (Attack)

- Real networks are resilient to random failures
- But, they are vulnerable in targeted attacks to high degree nodes

Error and attack tolerance of complex networks. R. Albert, H. Jeong and A.-L. Barabási. Nature, 2000

Response to Failures or Attacks



Cluster size distribution of SF network under attacks

Cluster size distribution of SF network under failures

f≈0.05 f≈0.18 f≈0.45

Fraction of removed nodes

Why Care about Modeling?

- Translate a physical problem to a mathematical one
  - Models can be used to prove properties – although they might not be the best ones
  - Better understanding of the data
- Anomaly detection
  - Find the distribution that a quantity should follow, and spot outliers (points that deviate from it)
- Answer what-if scenarios
  - What if somebody has twice as many friends? Do twice as many people adopt a product?

Next Lecture

- Part A:
  - More power-law distributions in real networks
  - Properties of time-evolving networks
  - Generative models for evolving networks
- Part B:
  - Centrality criteria and link analysis algorithms

Thank You!

http://www.travelandleisure.com/