Measuring Networks, and Random Graph Models

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

Network Properties: How to Measure a Network?

Plan: Key Network Properties

Degree distribution: $P(k)$

Path length: h

Clustering coefficient: C

Connected components: s

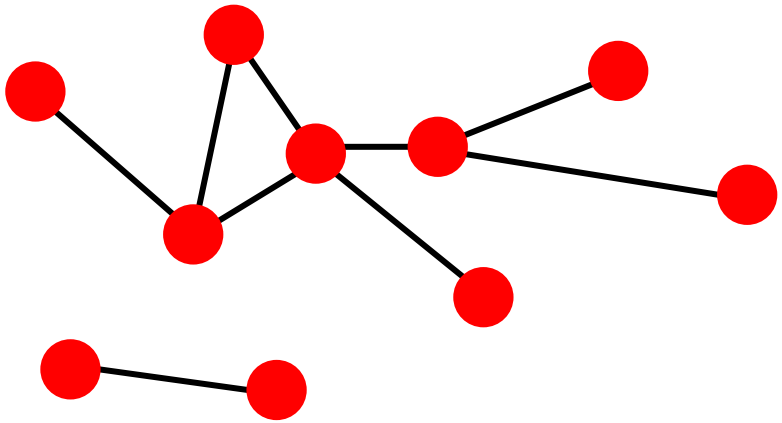
(1) Degree Distribution

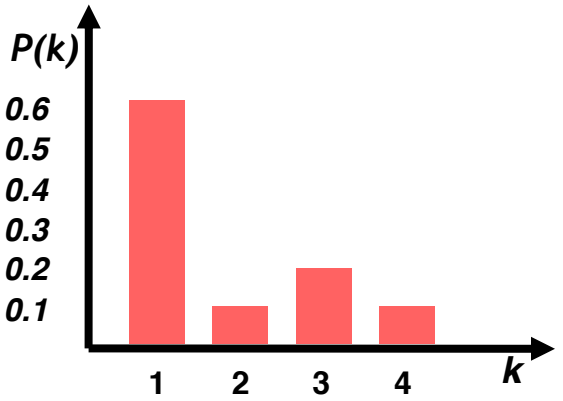
- **Degree distribution $P(k)$:** Probability that a randomly chosen node has degree k

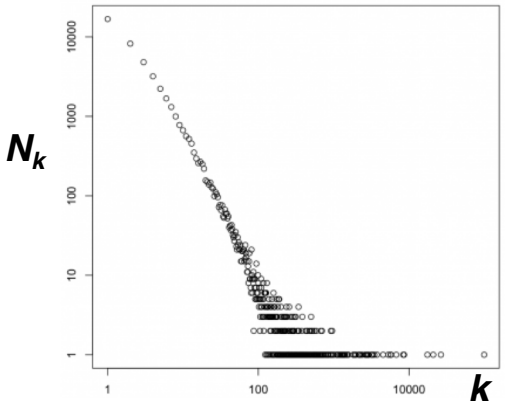
  N_k = # nodes with degree k

- Normalized histogram:

  $P(k) = N_k / N$ → **plot**





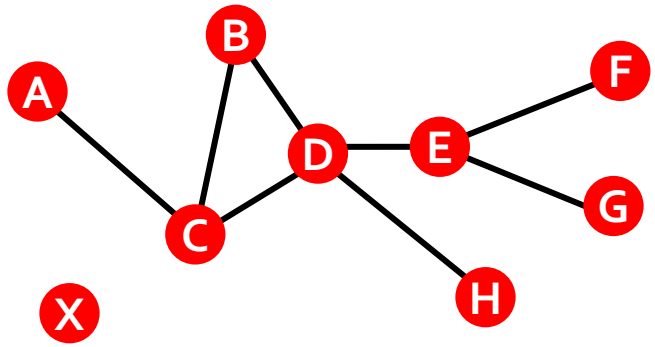


(2) Paths in a Graph

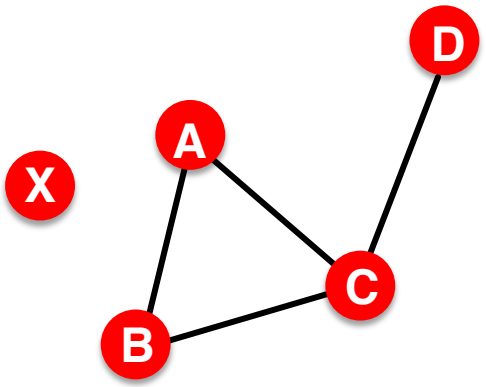
- A ***path*** is a sequence of nodes in which each node is linked to the next one

$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$

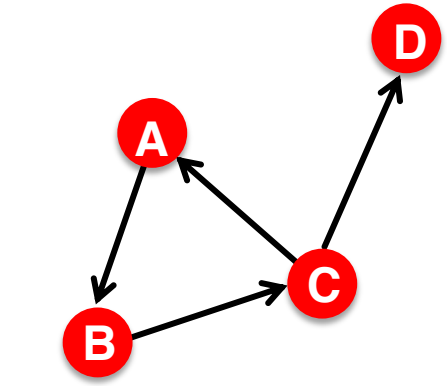
- Path can intersect itself and pass through the same edge multiple times
  - E.g.: ACBDCDEG
  - In a directed graph a path can only follow the direction of the "arrow"



Distance in a Graph



$h_{B,D} = 2$
$h_{A,X} = \infty$



$h_{B,C} = 1$, $h_{C,B} = 2$

- **Distance (shortest path, geodesic)** between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes
  - *If the two nodes are not connected, the distance is usually defined as infinite
- In **directed graphs** paths need to follow the direction of the arrows
  - Consequence: Distance is **not symmetric**: $h_{B,C} \neq h_{C,B}$

Network Diameter

- **Diameter:** The maximum (shortest path) distance between any pair of nodes in a graph

- **Average path length** for a connected graph (component) or a strongly connected (component of a) directed graph

$$\bar{h} = \frac{1}{2E_{\max}} \sum_{i,j \neq i} h_{ij}$$

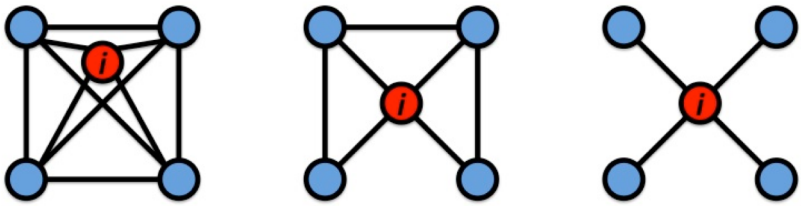
where h_{ij} is the distance from node i to node j
E_{max} is max number of edges (total number of node pairs) $= n(n-1)/2$

  - Many times we compute the average only over the connected pairs of nodes (that is, we ignore "infinite" length paths)

(3) Clustering Coefficient

- **Clustering coefficient:**
  - What portion of ***i***'s neighbors are connected?
  - Node ***i*** with degree k_i
  - $C_i \in [0,1]$
  - $C_i = \dfrac{2e_i}{k_i(k_i-1)}$

    where e_i is the number of edges between the neighbors of node i



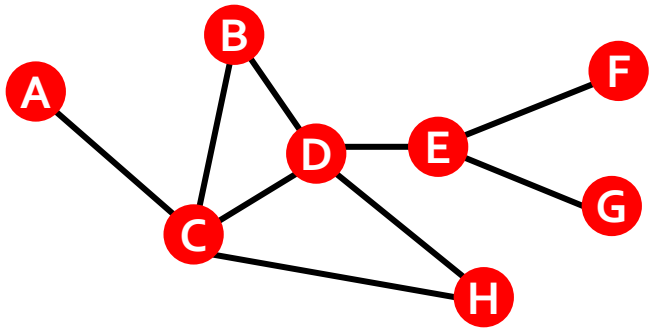
$C_i = 1$ $C_i = 1/2$ $C_i = 0$

- **Average clustering coefficient:** $C = \dfrac{1}{N}\sum_{i}^{N} C_i$

Clustering Coefficient

- **Clustering coefficient:**
  - What portion of ***i***'s neighbors are connected?
  - Node i with degree $\boldsymbol{k_i}$
  - $C_i = \dfrac{2e_i}{k_i(k_i-1)}$

where e_i is the number of edges between the neighbors of node i



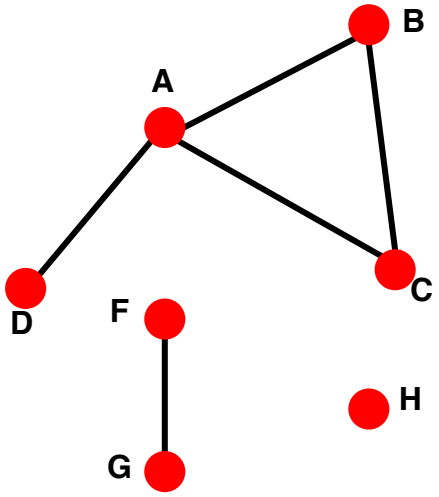
$k_B=2,\ e_B=1,\ C_B=2/2 = 1$

$k_D=4,\ e_D=2,\ C_D=4/12 = 1/3$

Avg. clustering: C=0.33

(4) Connectivity

- **Size of the largest connected component**
  - Largest set where any two vertices can be joined by a path
- **Largest component = Giant component**



How to find connected components:

- Start from random node and perform Breadth First Search (BFS)
- Label the nodes BFS visited
- If all nodes are visited, the network is connected
- Otherwise find an unvisited node and repeat BFS

Summary: Key Network Properties

Degree distribution: $P(k)$

Path length: h

Clustering coefficient: C

Connected components: s

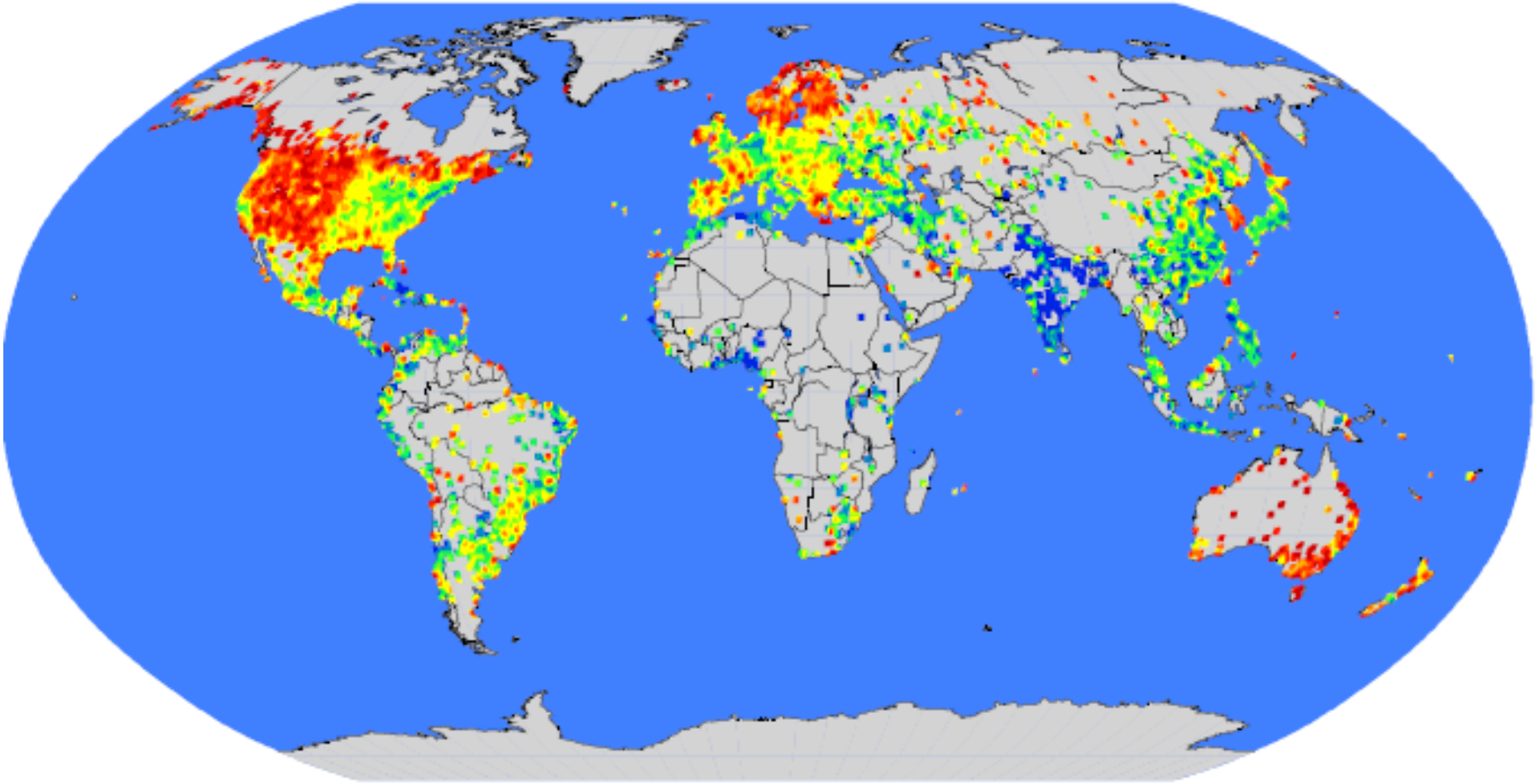
Let's measure P(k), h and C on a real-world network!

MSN Messenger

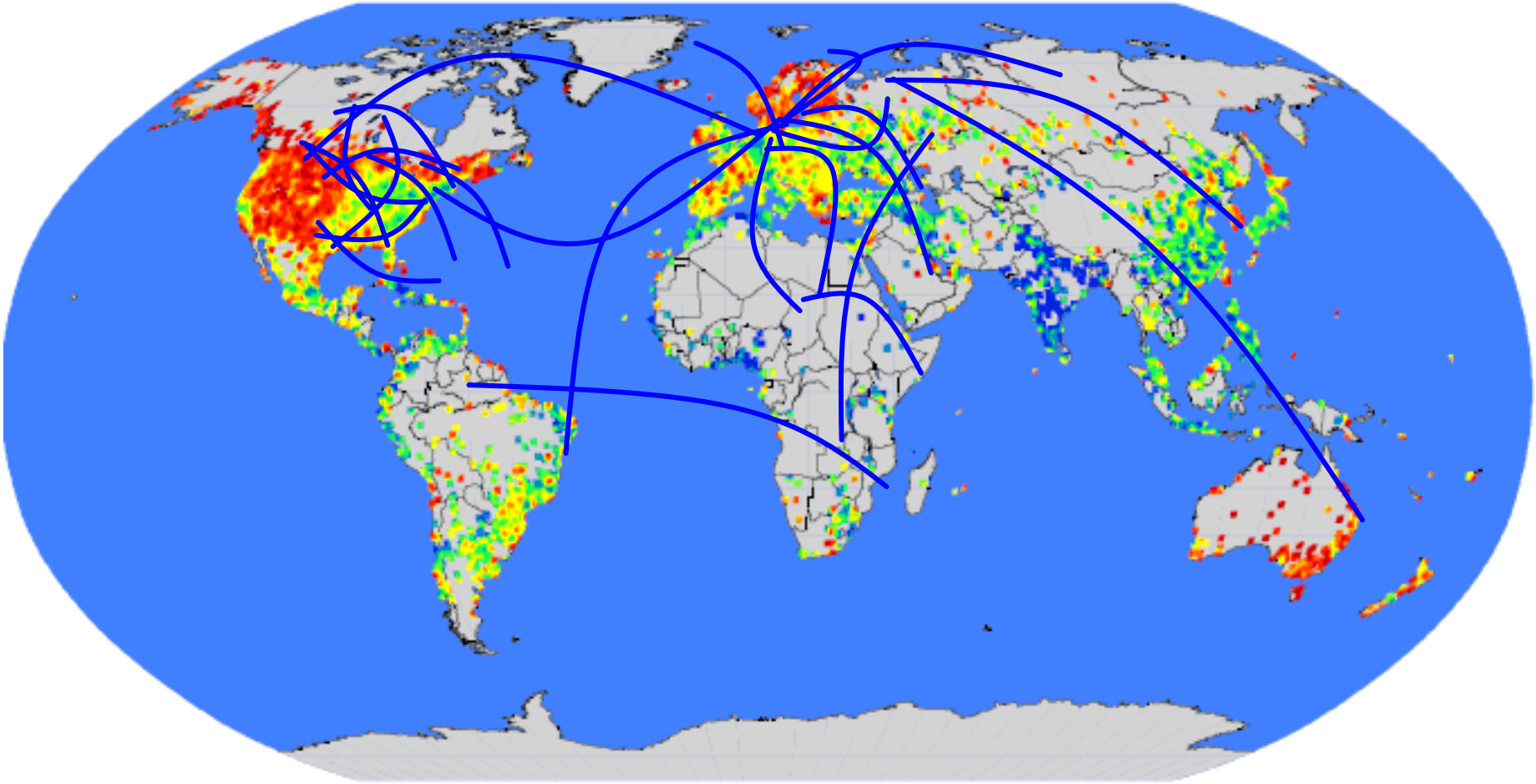
MSN Messenger.

- 1 month activity
  - 245 million users logged in
  - 180 million users engaged in conversations
  - More than 30 billion conversations
  - More than 255 billion exchanged messages

Communication: Geography

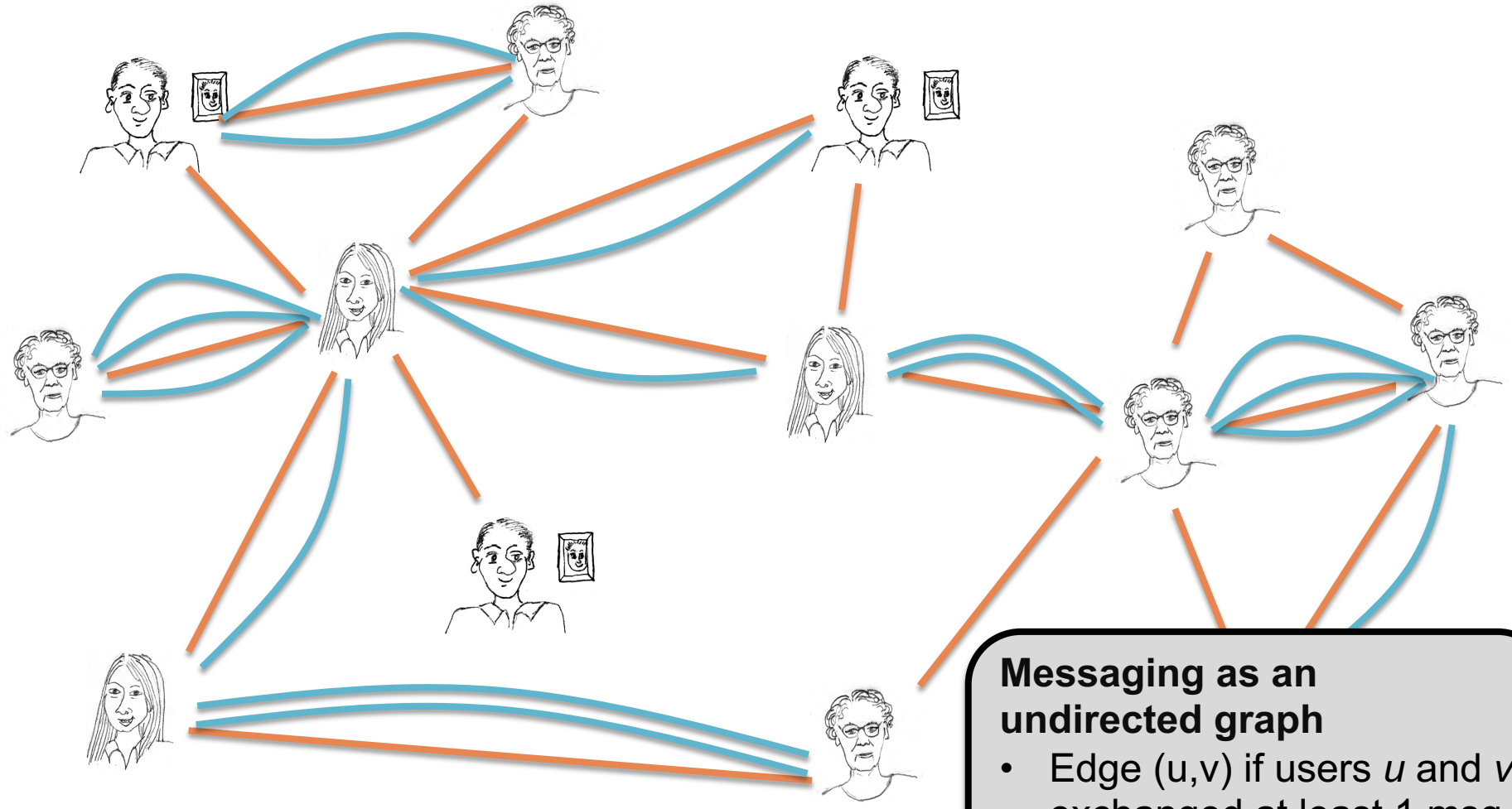


Communication Network



Network: 180M people, 1.3B edges

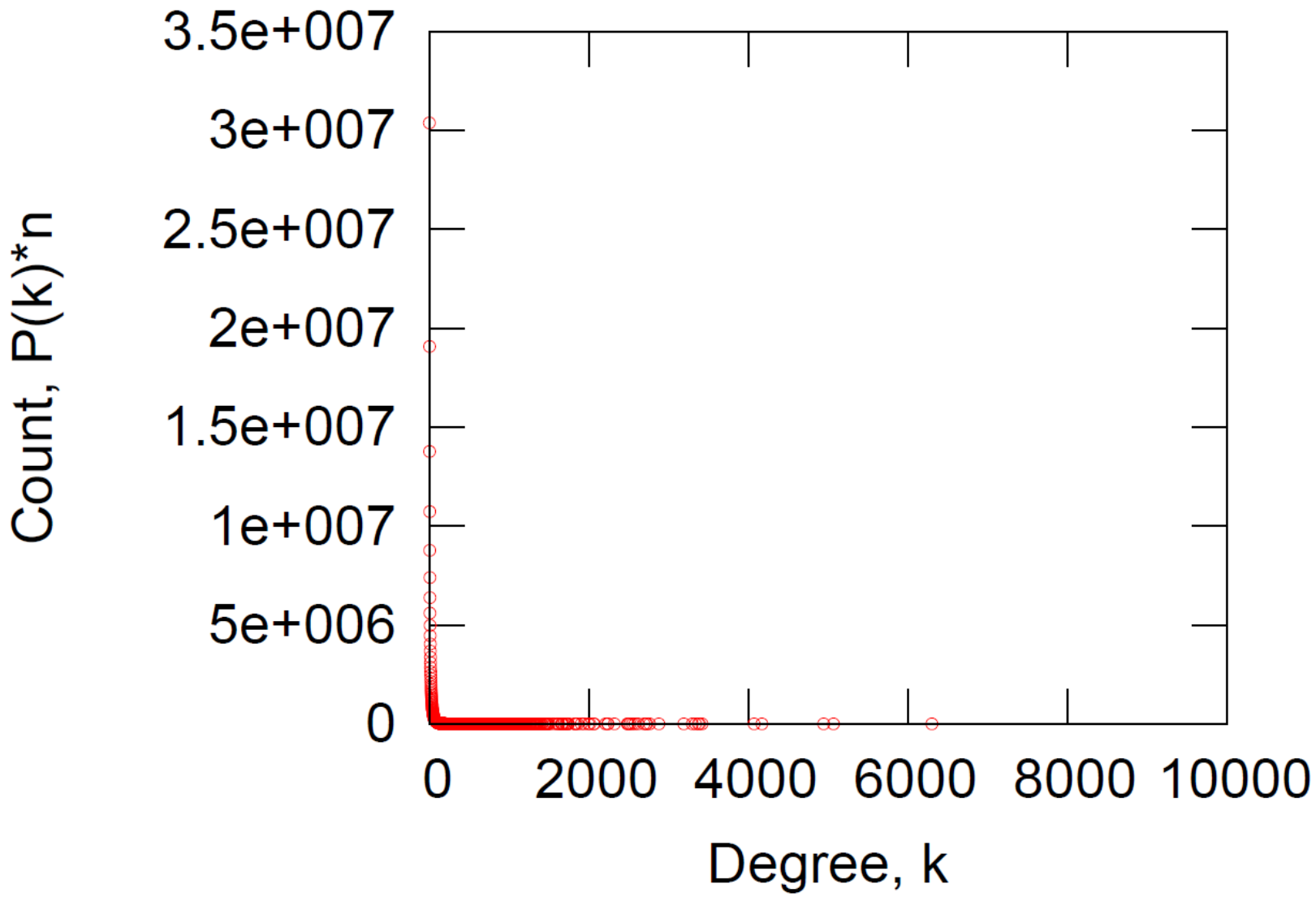
Messaging as a Multigraph



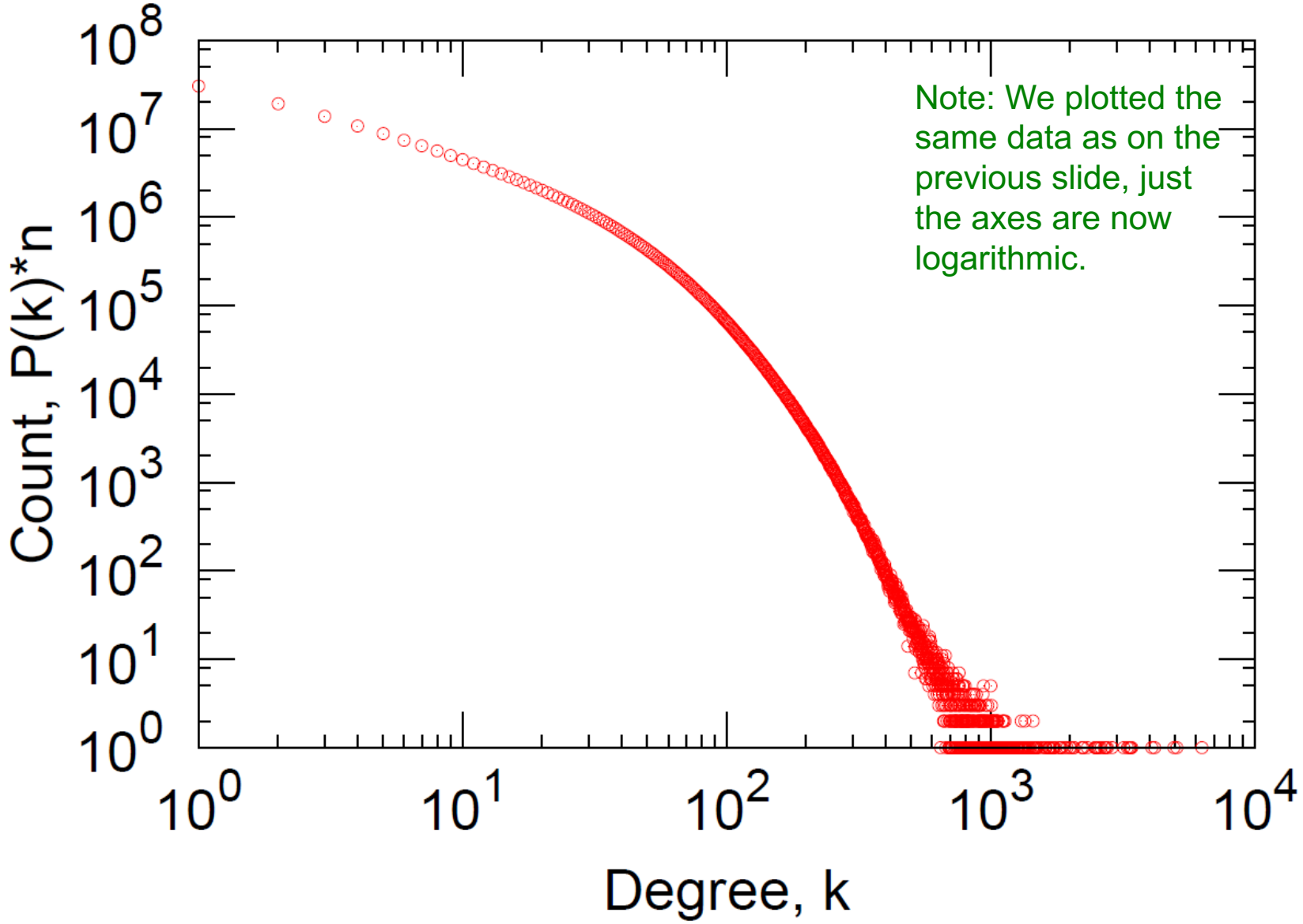
Messaging as an undirected graph

- Edge (u,v) if users *u* and *v* exchanged at least 1 msg
- N=180 million people
- E=1.3 billion edges

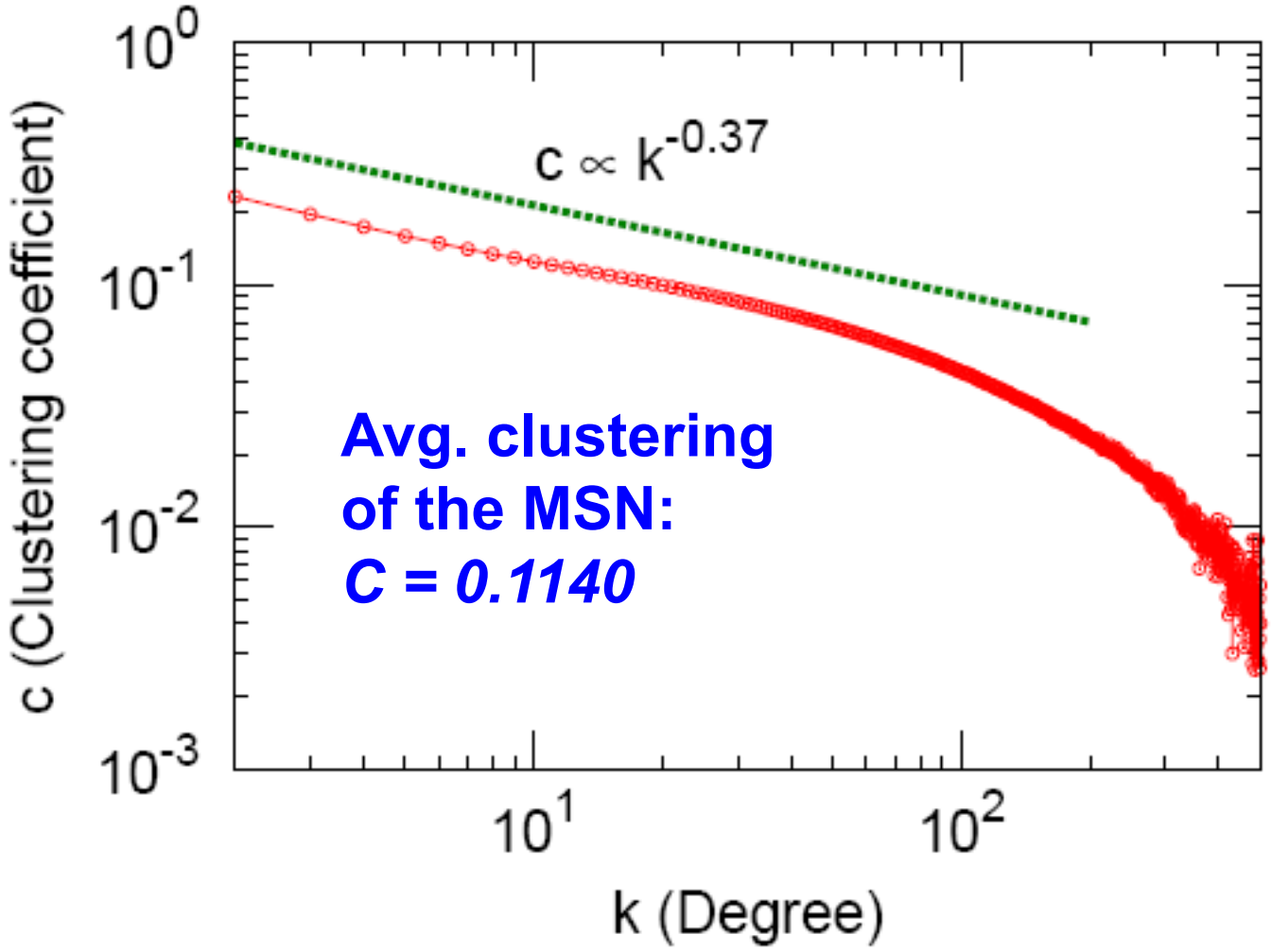
MSN: (1) Degree Distribution



MSN: Log-Log Degree Distribution

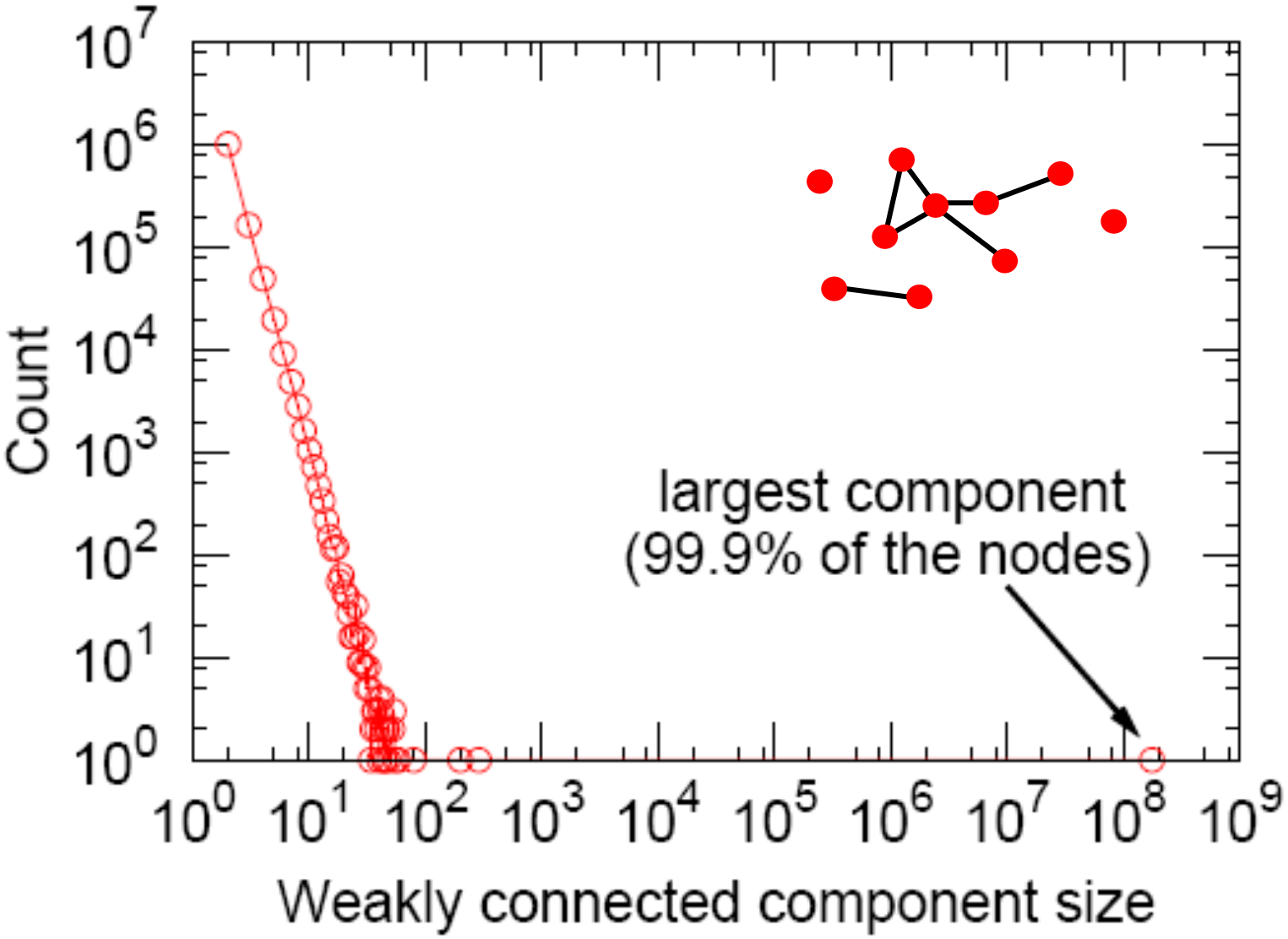


MSN: (2) Clustering

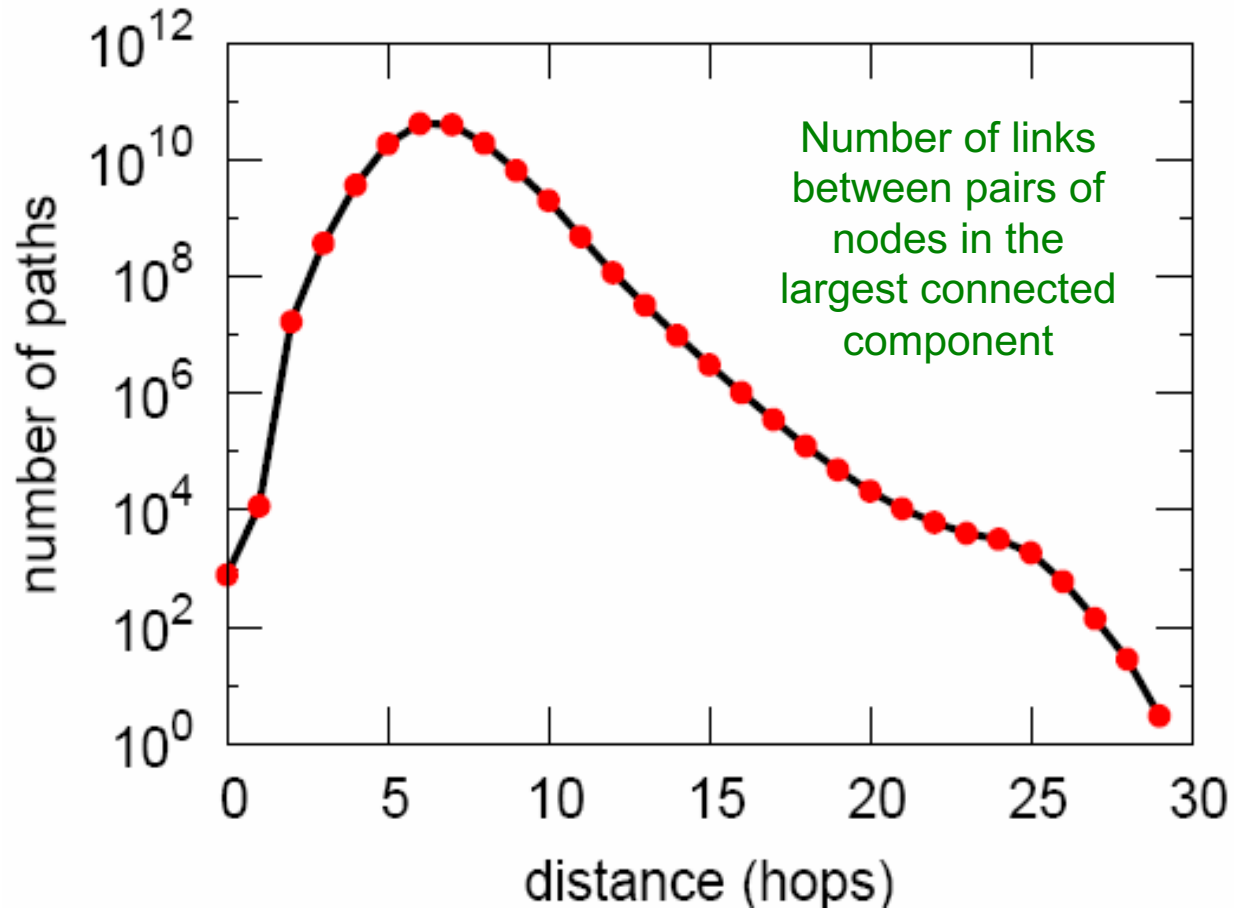


C_k: average C_i of nodes i of degree k: $C_k = \frac{1}{N_k} \sum_{i:k_i=k} C_i$

MSN: (3) Connected Components



MSN: (4) Diameter of WCC



Avg. path length **6.6**
90% of the nodes can be reached in < 8 hops

nodes as we do BFS out of a random node

Steps	#Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

MSN: Key Network Properties

Degree distribution: *Heavily skewed*
avg. degree= 14.4

Path length: *6.6*

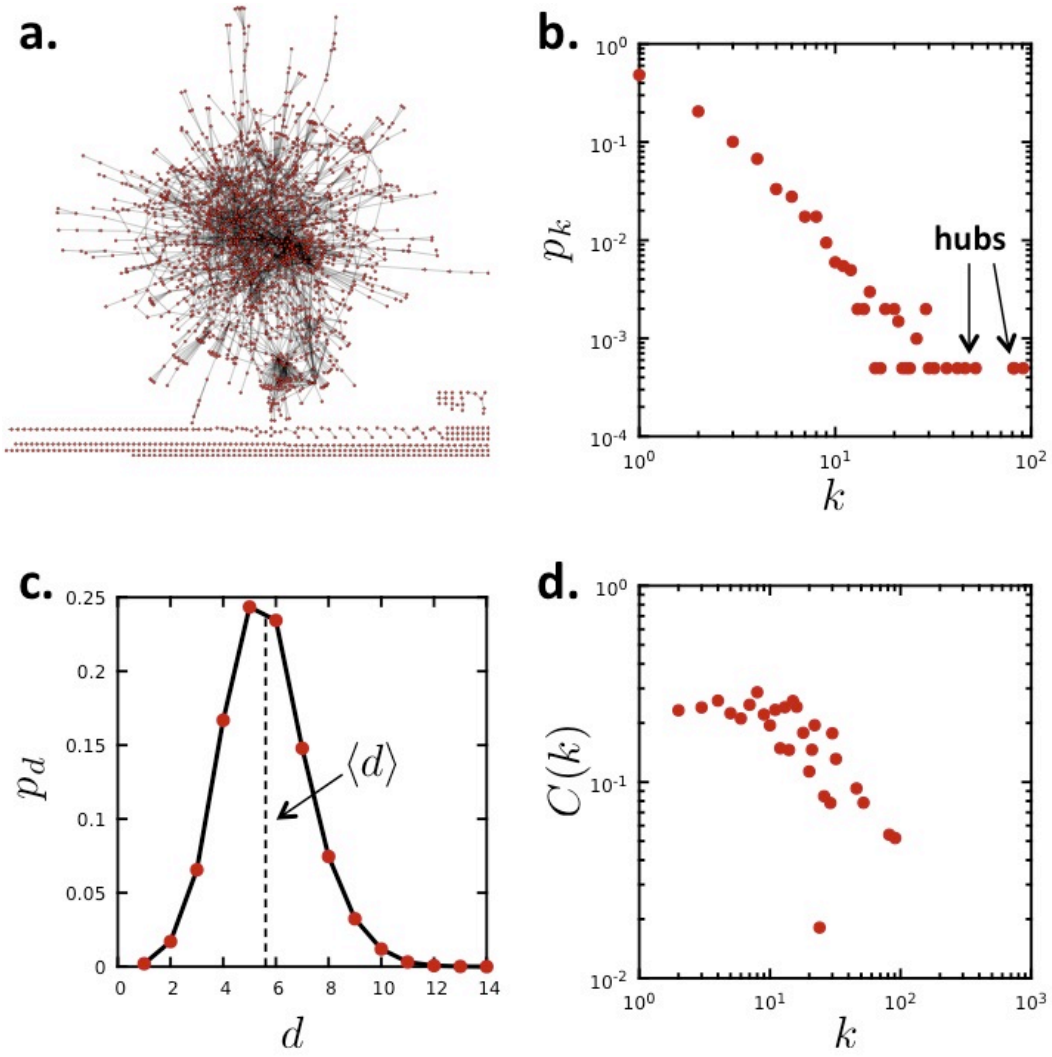
Clustering coefficient: *0.11*

Connectivity: *giant component*

Are these values "expected"?
Are they "surprising"?

To answer this we need a null-model!

Another example: PPI Network



a. Undirected network

N=2,018 proteins as nodes

E=2,930 binding interactions as links.

b. Degree distribution:

Skewed. Average degree <k>=2.90

c. Diameter:

Avg. path length = 5.8

d. Clustering:

Avg. clustering = 0.12

Connectivity: 185 components

the largest component 1,647 nodes (81% of nodes)

Erdös-Renyi Random Graph Model

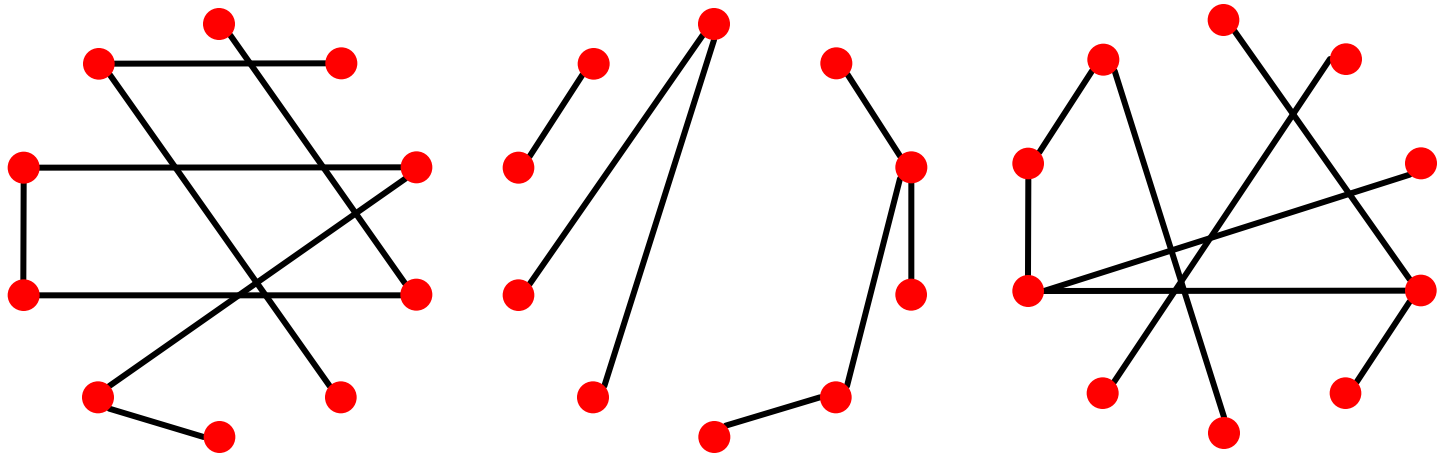
Simplest Model of Graphs

- **Erdös-Renyi Random Graphs** [Erdös-Renyi, '60]
- **Two variants:**
  - $G_{n,p}$: undirected graph on n nodes and each edge (u,v) appears i.i.d. with probability p
  - $G_{n,m}$: undirected graph with n nodes, and m uniformly at random picked edges

What kind of networks do such models produce?

Random Graph Model

- ***n*** **and** ***p*** **do not uniquely determine the graph!**
  - The graph is a result of a random process
- We can have many different realizations given the same ***n*** and ***p***



n = 10
p= 1/6

Properties of G_{np}

Degree distribution: $P(k)$

Path length: h

Clustering coefficient: C

What are the values of these properties for G_{np}?

Degree Distribution

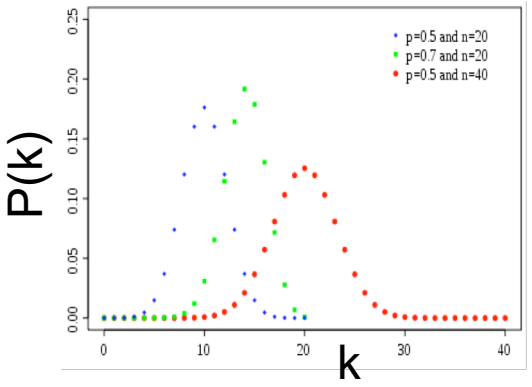
- **Fact:** Degree distribution of G_{np} is binomial.
- Let $P(k)$ denote the fraction of nodes with degree k:

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Select k nodes out of $n-1$

Probability of having k edges

Probability of missing the rest of the $n-1-k$ edges



Mean, variance of a binomial distribution

$$\bar{k} = p(n-1)$$

$$\sigma^2 = p(1-p)(n-1)$$

$$\frac{\sigma}{\bar{k}} = \left[\frac{1-p}{p}\frac{1}{(n-1)}\right]^{1/2} \approx \frac{1}{(n-1)^{1/2}}$$

By the law of large numbers, as the network size increases, the distribution becomes increasingly narrow—we are increasingly confident that the degree of a node is in the vicinity of k.

Clustering Coefficient of G_{np}

- **Remember:** $C_i = \frac{2e_i}{k_i(k_i-1)}$

  Where e_i is the number of edges between i's neighbors

- Edges in G_{np} appear i.i.d. with prob. p

- **So, expected E[e_i] is:** $= p\frac{k_i(k_i-1)}{2}$

  Each pair is connected with prob. p

  Number of distinct pairs of neighbors of node i of degree k_i

- **Then *E[C]*:** $= \frac{p \cdot k_i(k_i-1)}{k_i(k_i-1)} = p = \frac{\bar{k}}{n-1} \approx \frac{\bar{k}}{n}$

Clustering coefficient of a random graph is small.
If we generate bigger and bigger graphs with fixed avg. degree k (that is we set $p = k \cdot 1/n$), then *C* decreases with the graph size *n*.

Network Properties of G_{np}

Degree distribution: $P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$

Clustering coefficient: $C = p = \bar{k}/n$

Path length: ***next!***

Connectivity:

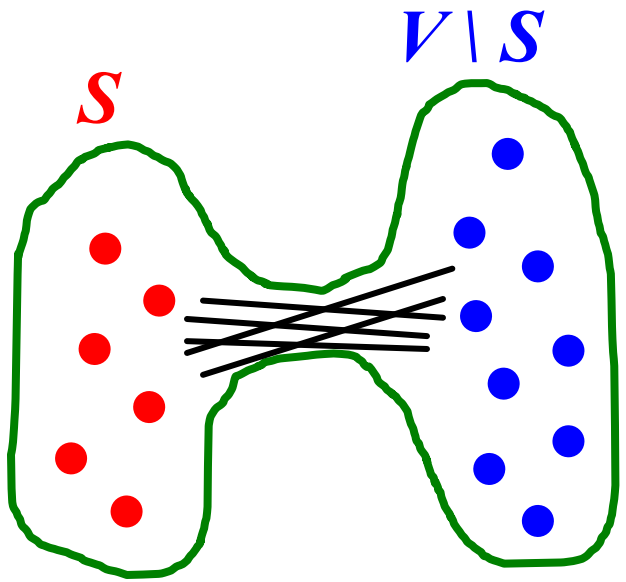
Def: Expansion

- Graph $G(V, E)$ has **expansion α**: if $\forall S \subseteq V$:

  $$\#\text{ of edges leaving } S \geq \alpha \cdot \min(|S|, |V \backslash S|)$$

- **Or equivalently:**

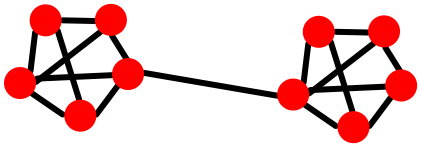
$$\alpha = \min_{S \subseteq V} \frac{\#edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$



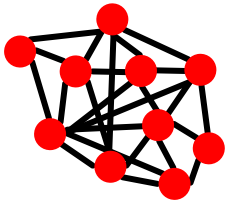
Expansion: Measures Robustness

$$\alpha = \min_{S \subseteq V} \frac{\#\,edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$

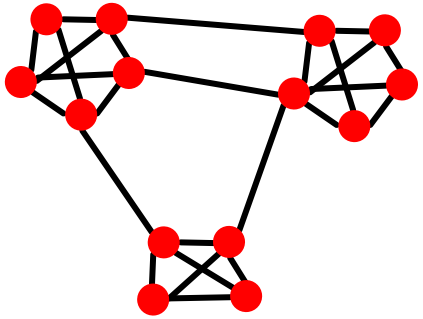
- Expansion is **measure of robustness:**
  - To disconnect l nodes, we need to cut $\geq \alpha \cdot l$ edges
- Low expansion:



- High expansion:

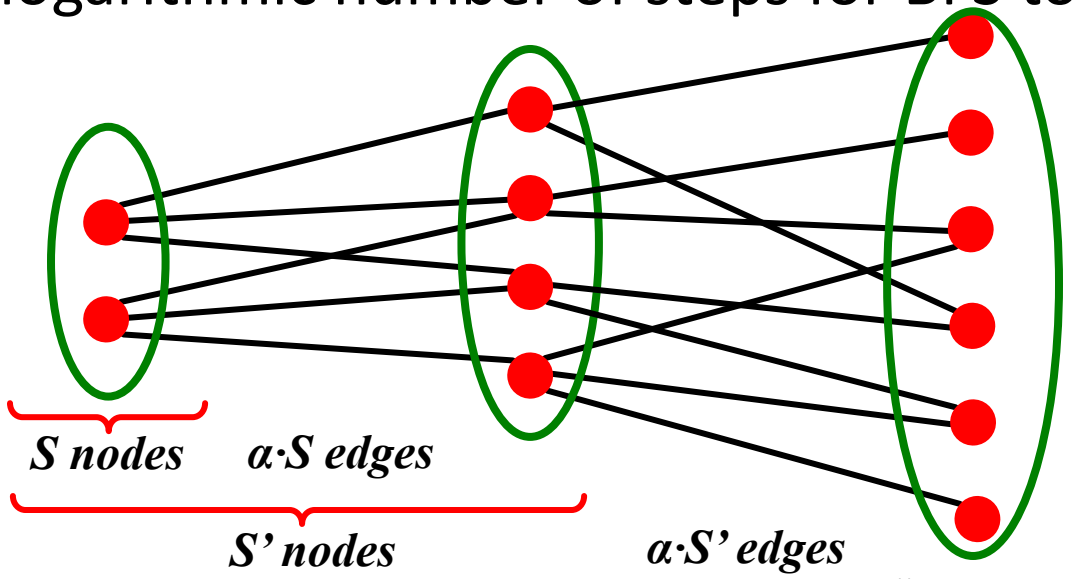


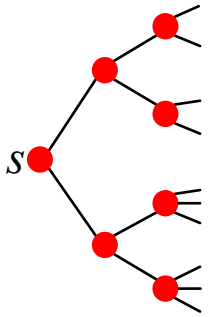
- Social networks:
  - "Communities"



Expansion: Random Graphs

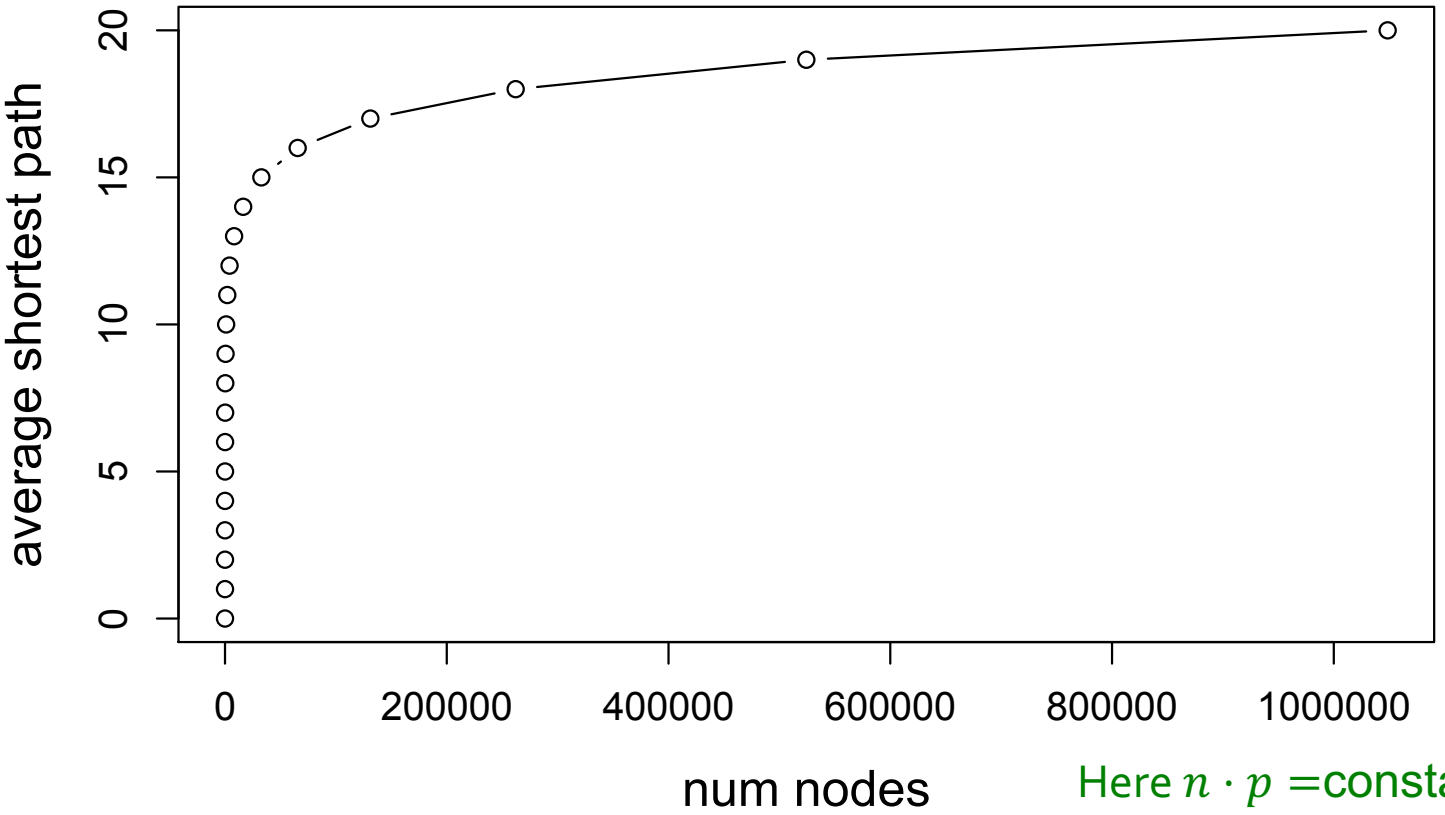
- **Fact:** In a graph on n nodes with expansion α for all pairs of nodes there is a path of length $O((\log n)/\alpha)$.
- **Random graph G_{np}:**
  For $\log n > np > c$, diam(G_{np}) = $O(\log n / \log (np))$
  - Random graphs have good expansion so it takes a logarithmic number of steps for BFS to visit all nodes





Erdös-Renyi avg. shortest path

Erdös-Renyi Random Graph can grow very large but nodes will be just a few hops apart



Here $n \cdot p$ =constant
That is, avg deg k is const

Network Properties of G_{np}

Degree distribution: $P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$

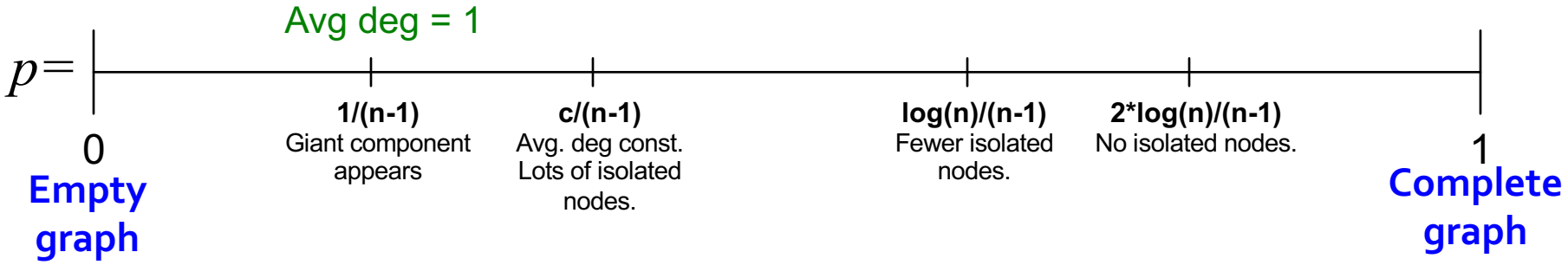
Path length: $O(\log n)$

Clustering coefficient: $C = p = \bar{k} / n$

Connected components: ***next!***

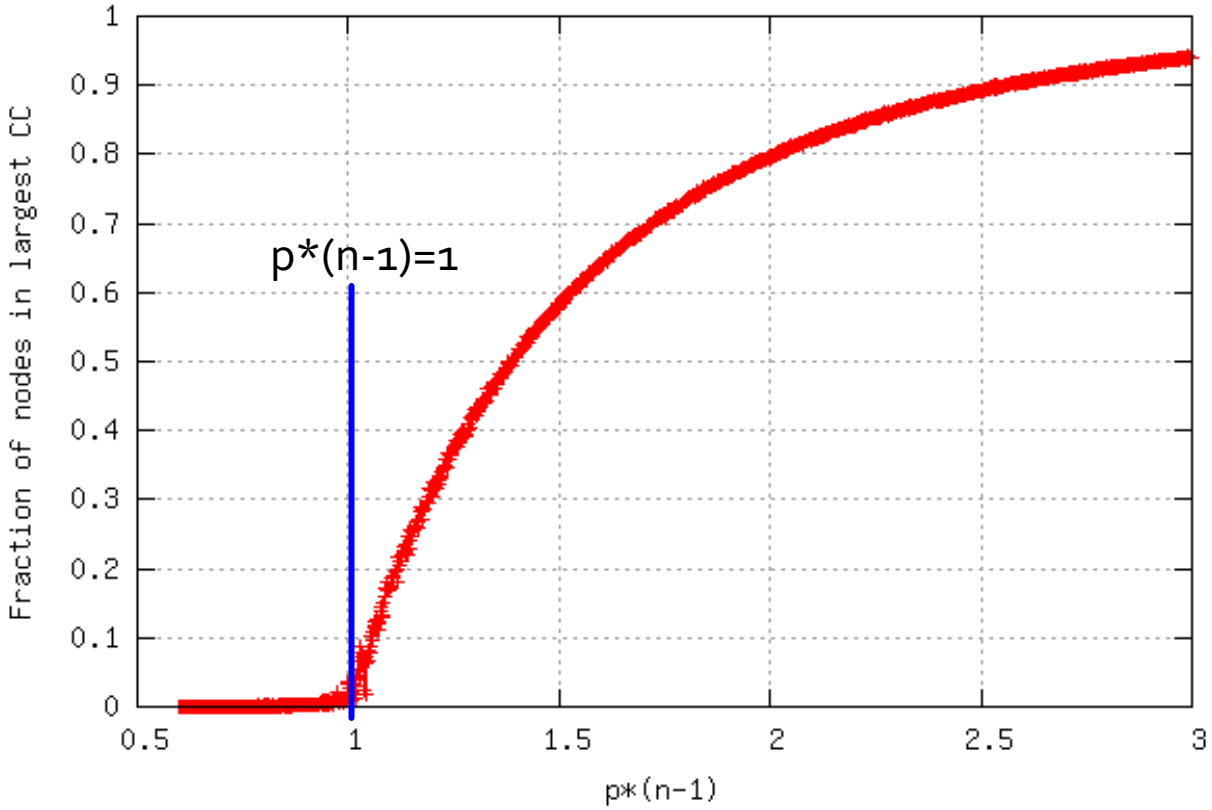
"Evolution" of a Random Graph

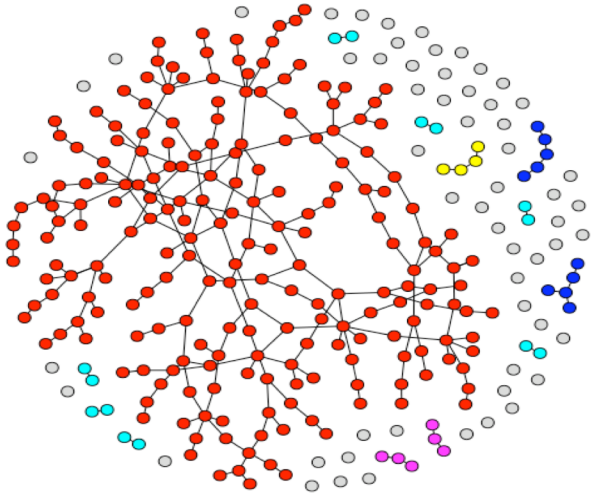
- **Graph structure of G_{np} as p changes:**



- **Emergence of a giant component:**
  avg. degree $k=2E/n$ or $p=k/(n-1)$
  - $k=1-\varepsilon$: all components are of size $\Omega(\log n)$
  - $k=1+\varepsilon$: 1 component of size $\Omega(n)$, others have size $\Omega(\log n)$
    - Each node has at least one edge in expectation

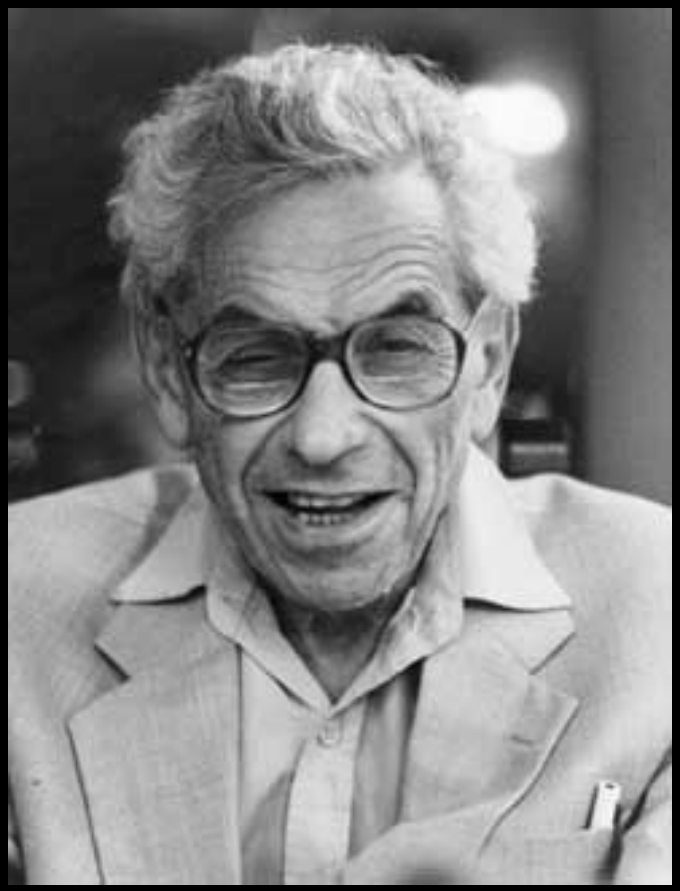
G_{np} Simulation Experiment





Fraction of nodes in the largest component

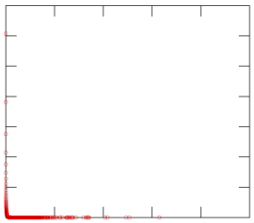
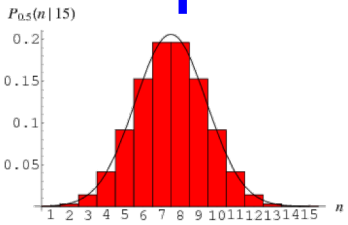
- G_{np}, n=100,000, $k=p(n-1)$ = 0.5 … 3



Paul Erdös

G_{np} is so cool!
Let's compare it to real networks.

Back to MSN vs. G_{np}

	MSN	G_{np} n=180M	
Degree distribution:			☒
Avg. path length:	**6.6**	$O(\log n)$ *h ≈ 8.2*	☑
Avg. clustering coef.:	*0.11*	$\bar{k}/n$ *C ≈ 8·10⁻⁸*	☒
Largest Conn. Comp.:	*99%*	GCC exists when $\bar{k}>1$. $\bar{k} \approx 14$.	☑

Real Networks vs. G_{np}

- **Are real networks like random graphs?**
  - Giant connected component: ☺
  - Average path length: ☺
  - Clustering Coefficient: ☹
  - Degree Distribution: ☹
- **Problems with the random networks model:**
  - Degree distribution differs from that of real networks
  - Giant component in most real network does NOT emerge through a phase transition
  - No local structure – clustering coefficient is too low
- **Most important: Are real networks random?**
  - The answer is simply: **NO!**

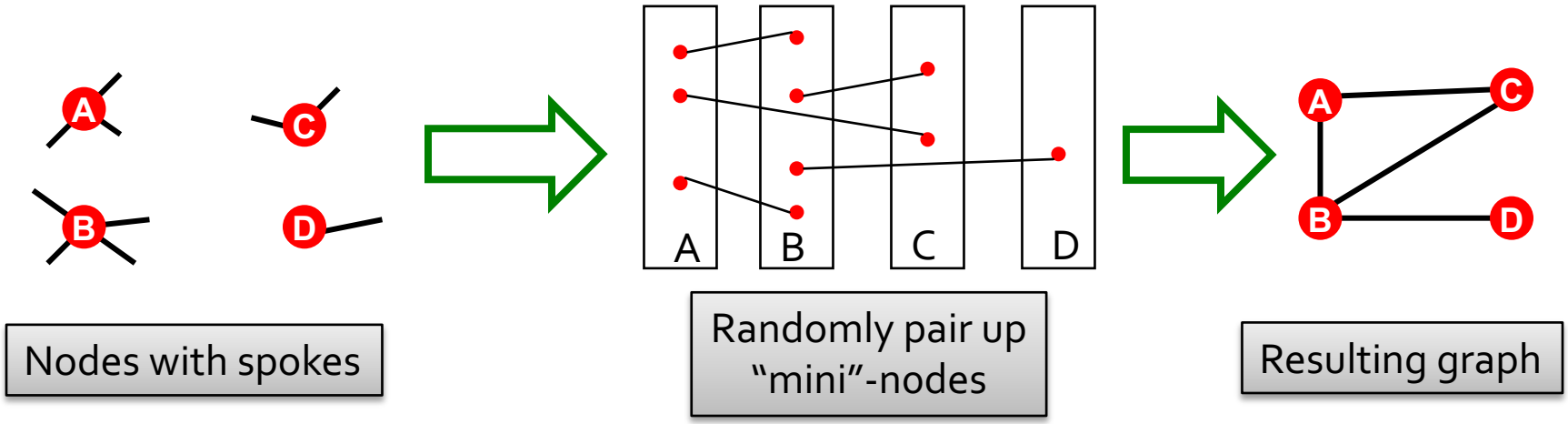
Real Networks vs. G_{np}

- **If G_{np} is wrong, why did we spend time on it?**
  - It is the reference model for the rest of the class
  - It will help us calculate many quantities, that can then be compared to the real data
  - It will help us understand to what degree is a particular property the result of some random process

So, while G_{np} is WRONG, it will turn out to be extremly USEFUL!

Intermezzo: Configuration Model

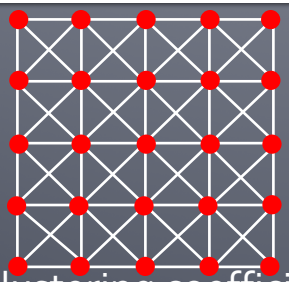
- **Goal:** Generate a random graph with a given degree sequence $k_1, k_2, \ldots k_N$
- **Configuration model:**



- **Useful as a "null" model of networks:**
  - We can compare the real network G and a "random" G' which has the same degree sequence as G

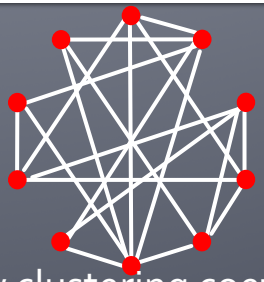
The Small-World Model

Can we have high clustering while also having short paths?



High clustering coefficient,
High diameter

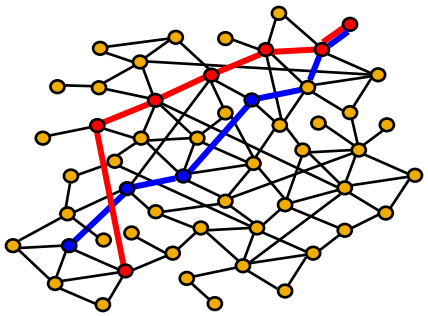
Vs.



Low clustering coefficient
Low diameter

The Small-World Experiment

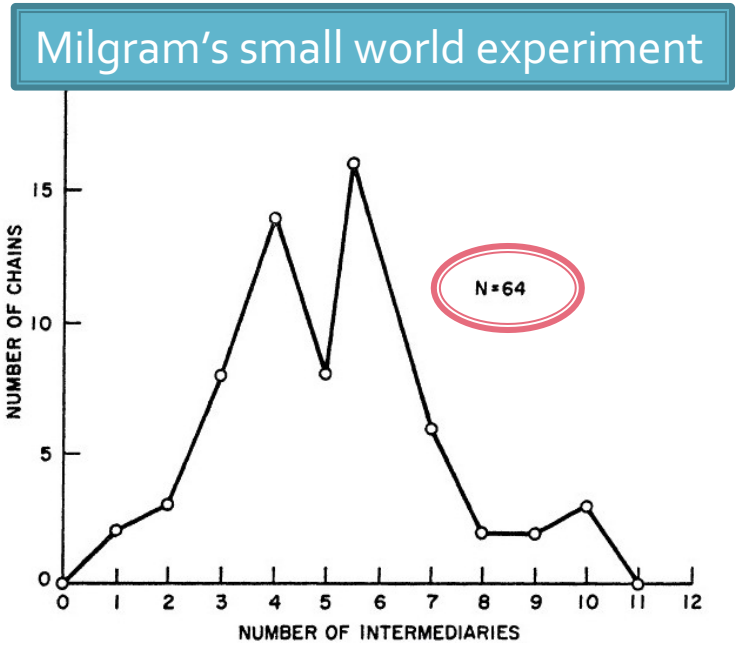
- **What is the typical shortest path length between any two people?**
  - **Experiment on the global friendship network**
    - Can't measure, need to probe explicitly
- **Small-world experiment** [Milgram '67]
  - Picked 300 people in Omaha, Nebraska and Wichita, Kansas
  - Ask them to get a letter to a stock-broker in Boston by passing it through friends
- **How many steps did it take?**



The Small-World Experiment

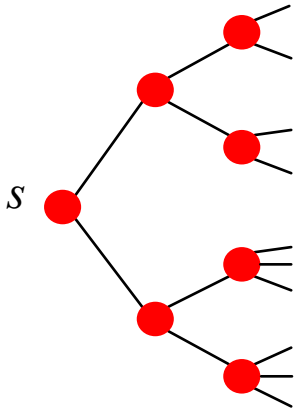
[Milgram, '67]

- **64 chains completed:**
  (i.e., 64 letters reached the target)
  - It took 6.2 steps on the average, thus
    "6 degrees of separation"
- **Further observations:**
  - People who owned stock had shorter paths to the stockbroker than random people: 5.4 vs. 6.7
  - People from the Boston area have even closer paths: 4.4

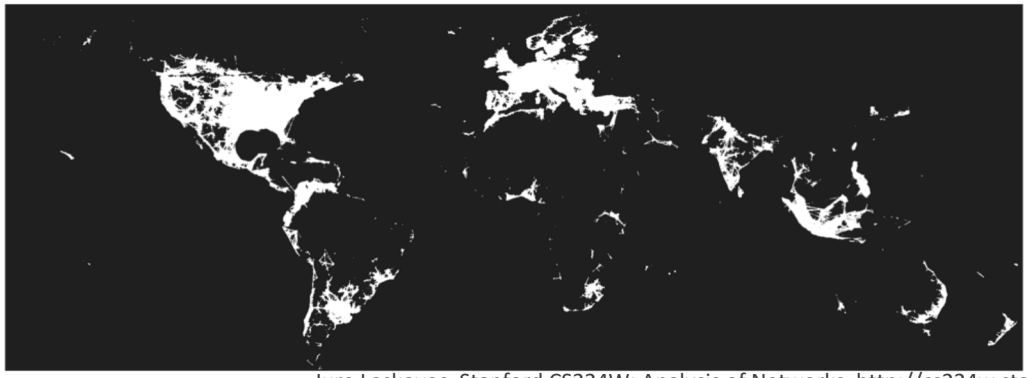


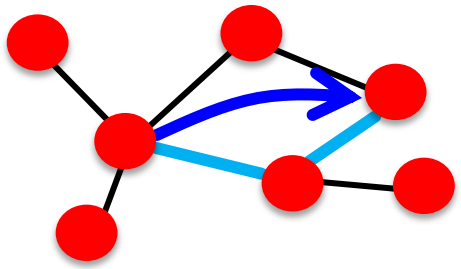
6-Degrees: Should We Be Surprised?

- **Assume each human is connected to 100 other people**
  Then:
  - Step 1: reach 100 people
  - Step 2: reach 100*100 = 10,000 people
  - Step 3: reach 100*100*100 = 1,000,000 people
  - Step 4: reach 100*100*100*100 = 100M people
  - **In 5 steps we can reach 10 billion people**



- **What's wrong here? We ignore clustering!**
  - **Not all edges point to new people**
    - **92% of FB friendships happen through a friend-of-a-friend**





Clustering Implies Edge Locality

- **MSN network has 7 orders of magnitude larger clustering than the corresponding G_{np}!**
- **Other examples:**

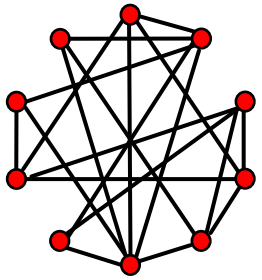
Actor Collaborations (IMDB): N = 225,226 nodes, avg. degree $\overline{k}$ = 61
Electrical power grid: N = 4,941 nodes, $\overline{k}$ = 2.67
Network of neurons: N = 282 nodes, $\overline{k}$ = 14

Network	h_{actual}	h_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power Grid	18.70	12.40	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

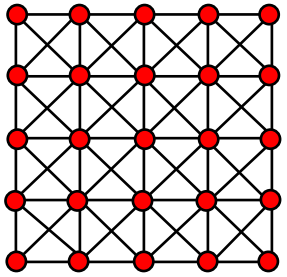
h ... Average shortest path length
C ... Average clustering coefficient
"actual" … real network
"random" … random graph with same avg. degree

The "Controversy"

- **Consequence of expansion:**
  - **Short paths: $O(\log n)$**
    - This is the smallest diameter we can get if we have a constant degree.
  - But clustering is low!
- **But networks have "local" structure:**
  - **Triadic closure:**
    Friend of a friend is my friend
  - High clustering but diameter is also high
- **How can we have both?**



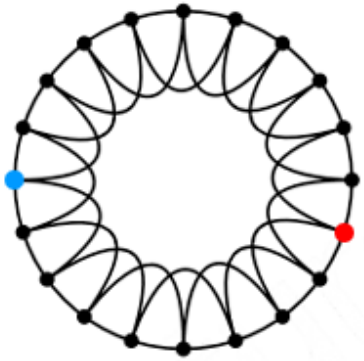
Low diameter
Low clustering coefficient



High clustering coefficient
High diameter

Small-World: How?

- **Could a network with high clustering also be a small world ($\log n$ dimeter)?**
  - How can we at the same time have **high clustering** and **small diameter?**



High clustering
High diameter

Low clustering
Low diameter

  - Clustering implies edge "locality"
  - Randomness enables "shortcuts"

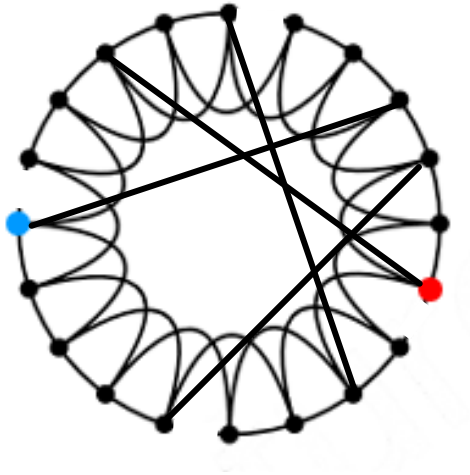
[Watts-Strogatz, '98]

Solution: The Small-World Model

Small-World Model [Watts-Strogatz '98]

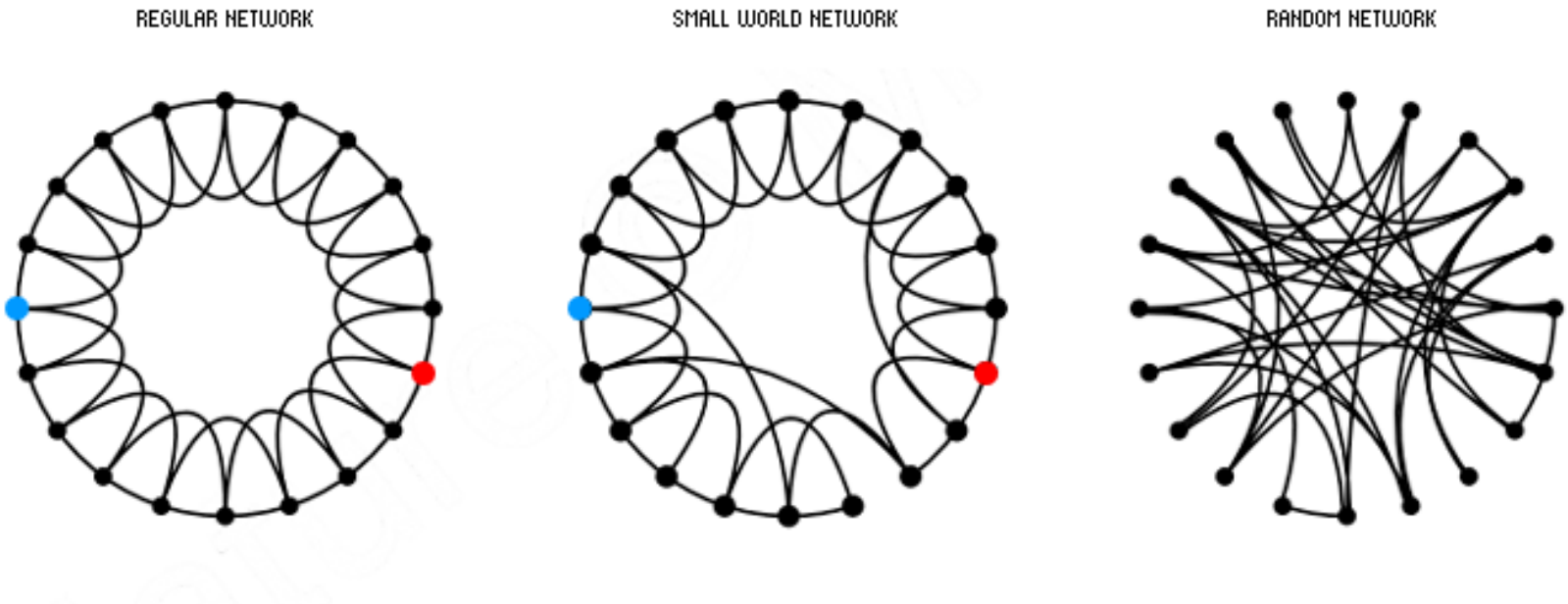
Two components to the model:

- **(1) Start with a low-dimensional regular lattice**
  - (In our case we are using a ring as a lattice)
  - Has high clustering coefficient
- **Now introduce randomness ("shortcuts")**
- **(2) Rewire:**
  - Add/remove edges to create shortcuts to join remote parts of the lattice
  - For each edge with prob. p move the other end to a random node



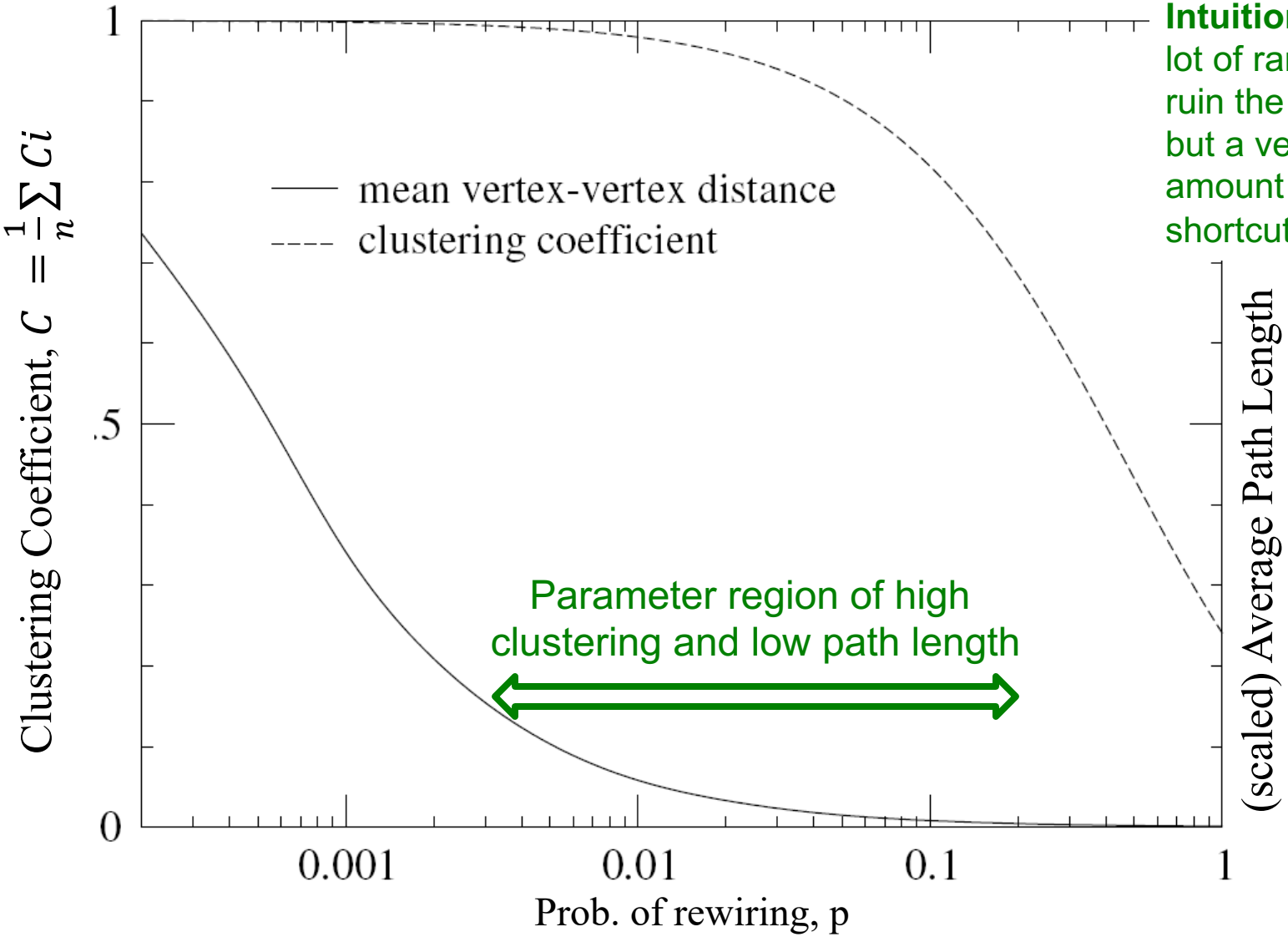
The Small-World Model

[Watts-Strogatz, '98]



Rewiring allows us to "interpolate" between a regular lattice and a random graph

The Small-World Model



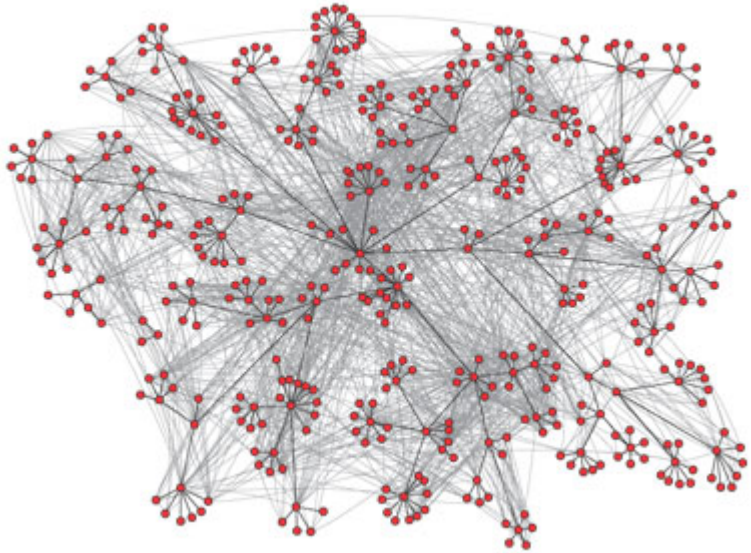
Intuition: It takes a lot of randomness to ruin the clustering, but a very small amount to create shortcuts.

Small-World: Summary

- **Could a network with high clustering be at the same time a small world?**
  - Yes! You don't need more than a few random links
- **The Watts Strogatz Model:**
  - Provides insight on the interplay between clustering and the small-world
  - Captures the structure of many realistic networks
  - Accounts for the high clustering of real networks
  - Does not lead to the correct degree distribution

Network Formation Processes

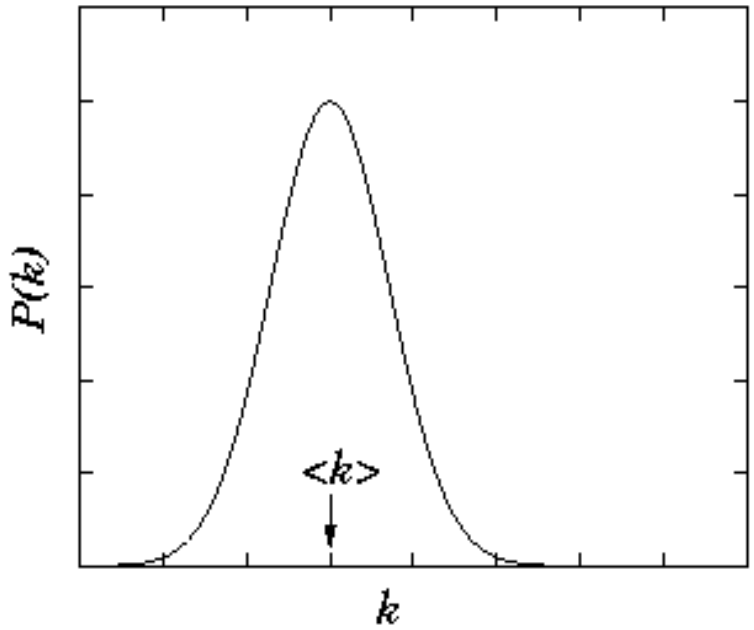
Which interesting graph properties do we observe that need explaining?



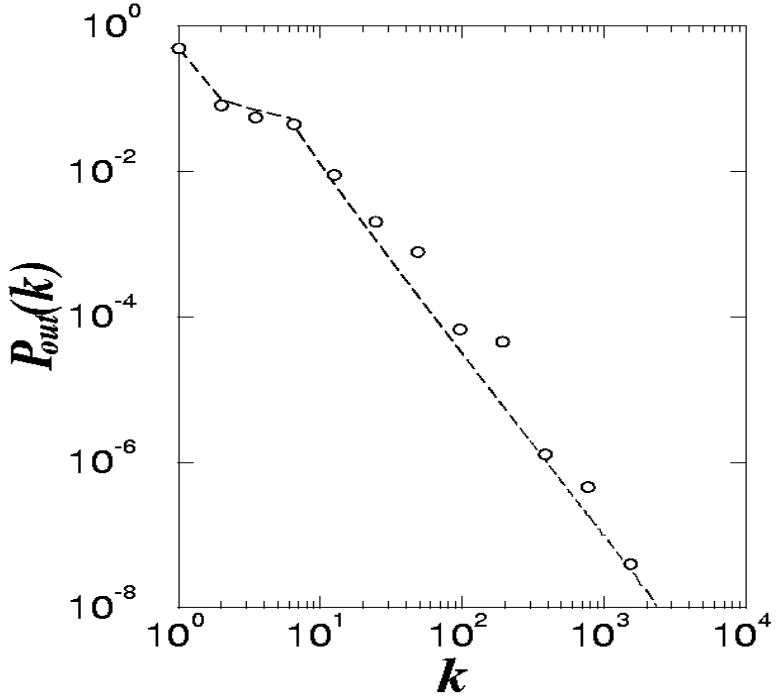
- **Small-world model:**
  - Diameter
  - Clustering coefficient
- **Node degree distribution**
  - What fraction of nodes has degree k (as a function of k)?
  - Prediction from simple random graph models: $p(k) =$ **exponential function of** k
  - **Observation: Often a power-law:** $p(k) \propto k^{-\alpha}$

Degree Distributions

Expected based on G_{np}



Found in data

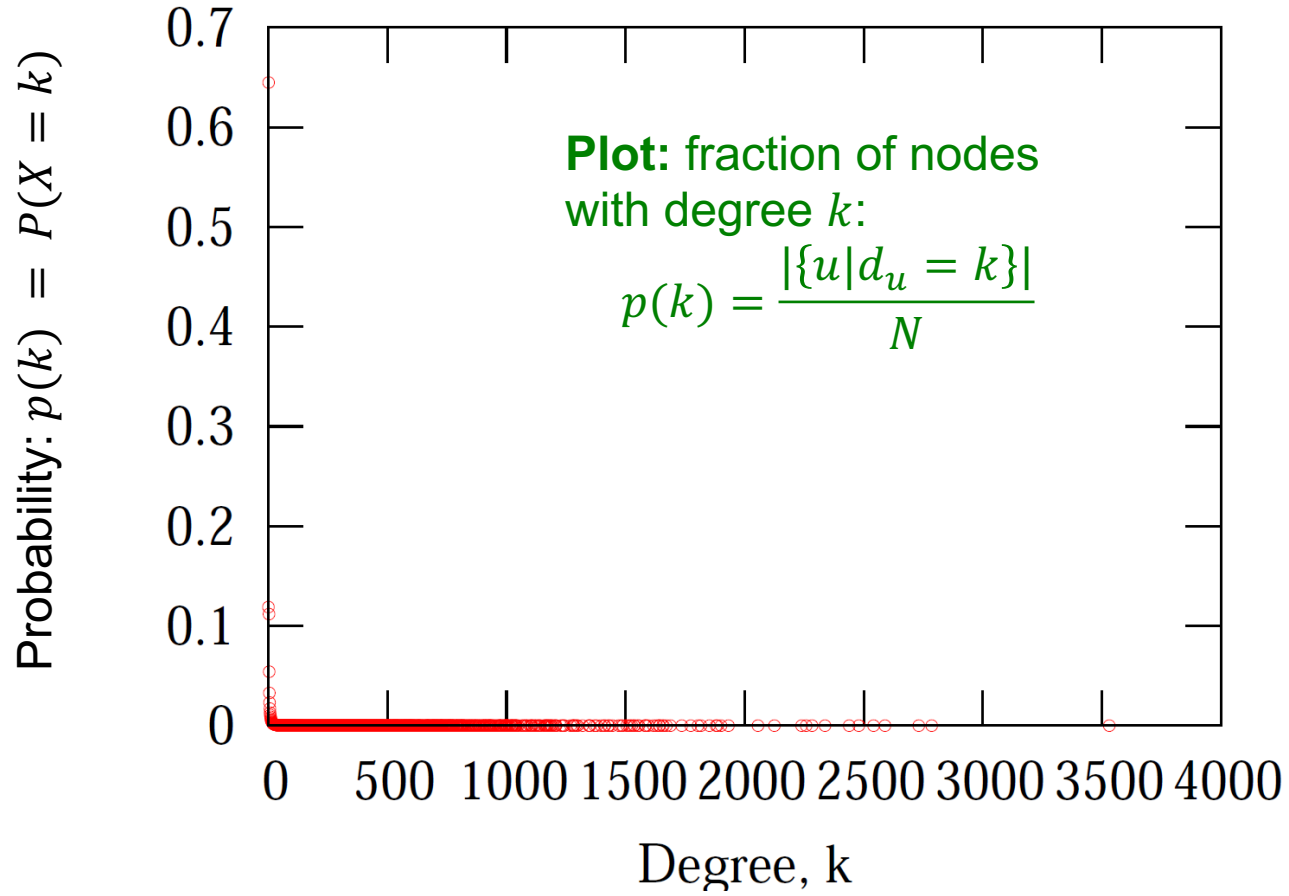


$$P(k) \propto k^{-\alpha}$$

[Leskovec et al. KDD '08]

Node Degrees in Networks

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

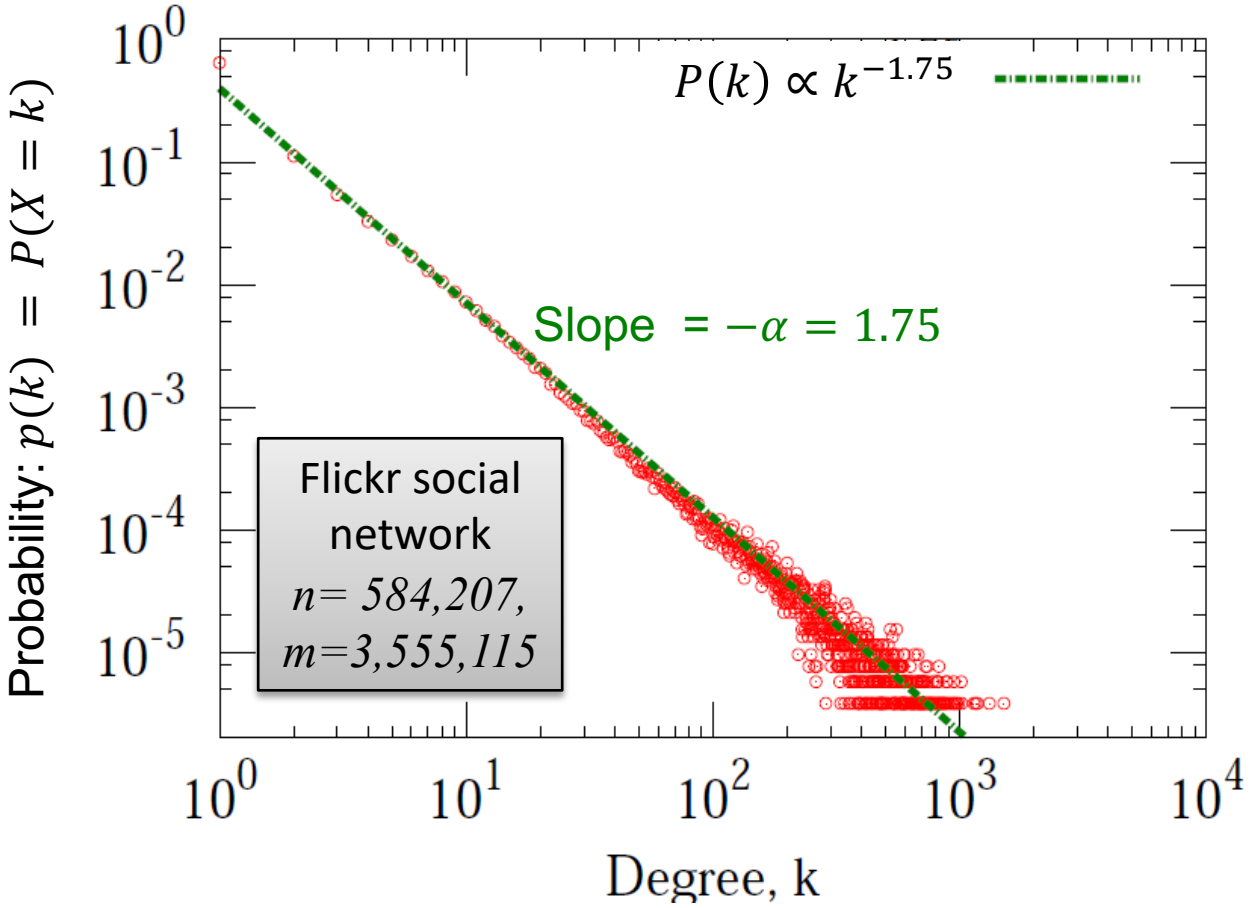


Flickr social network
n= 584,207, m=3,555,115

[Leskovec et al. KDD '08]

Node Degrees in Networks

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

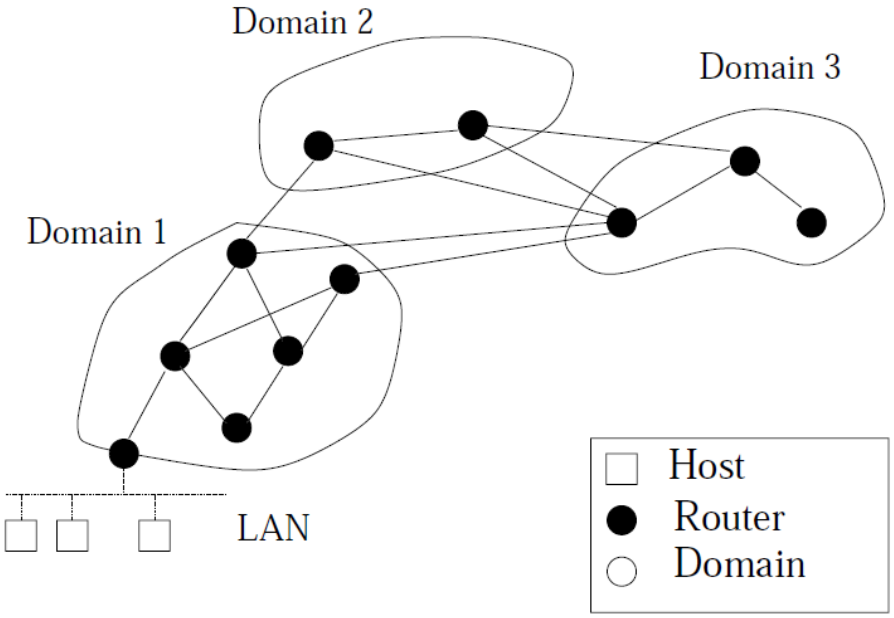


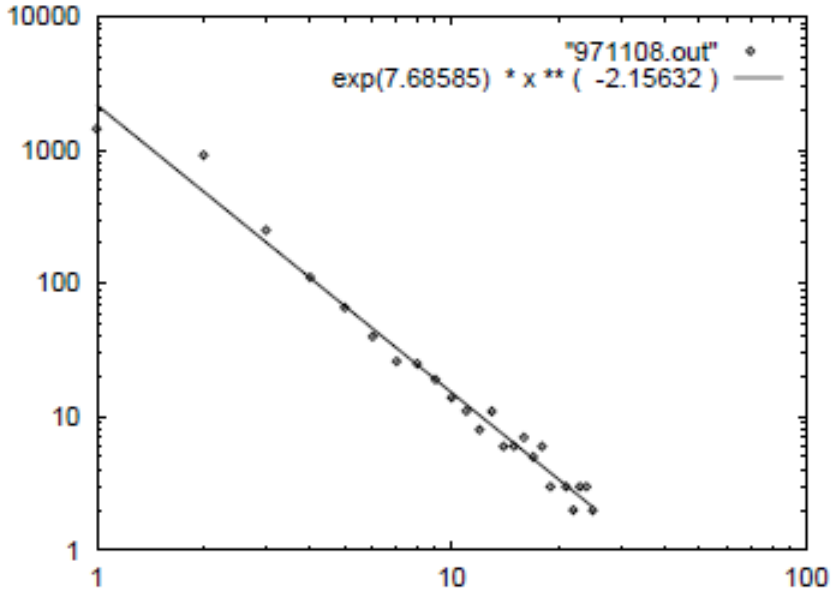
How to distinguish:
$P(k) \propto \exp(-k)$ **vs.**
$P(k) \propto k^{-\alpha}$ **?**

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]

- **First observed in Internet Autonomous Systems** [Faloutsos, Faloutsos and Faloutsos, 1999]

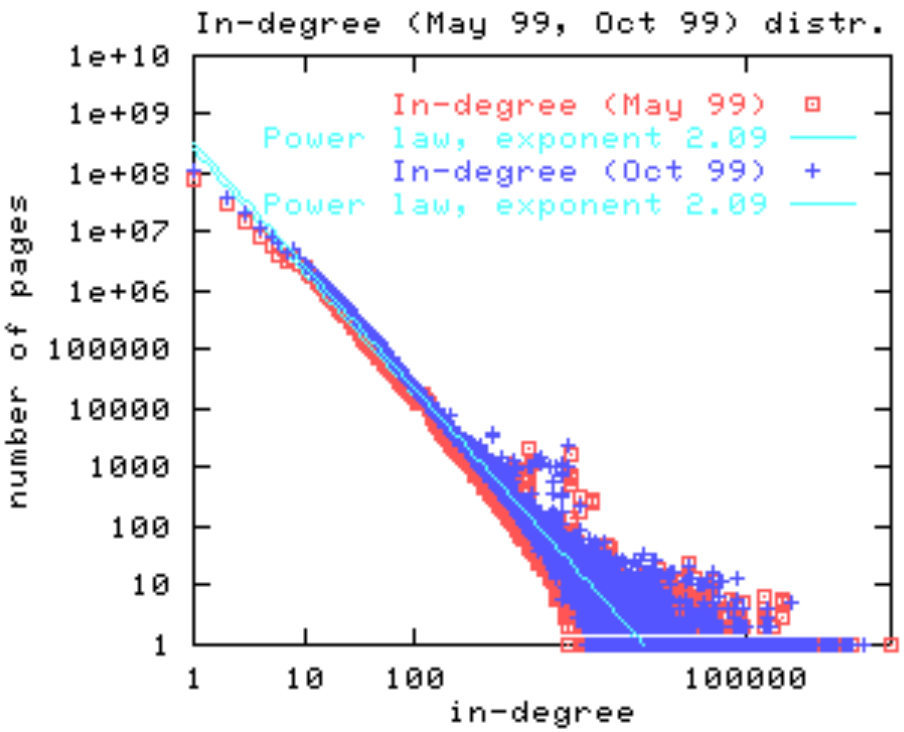


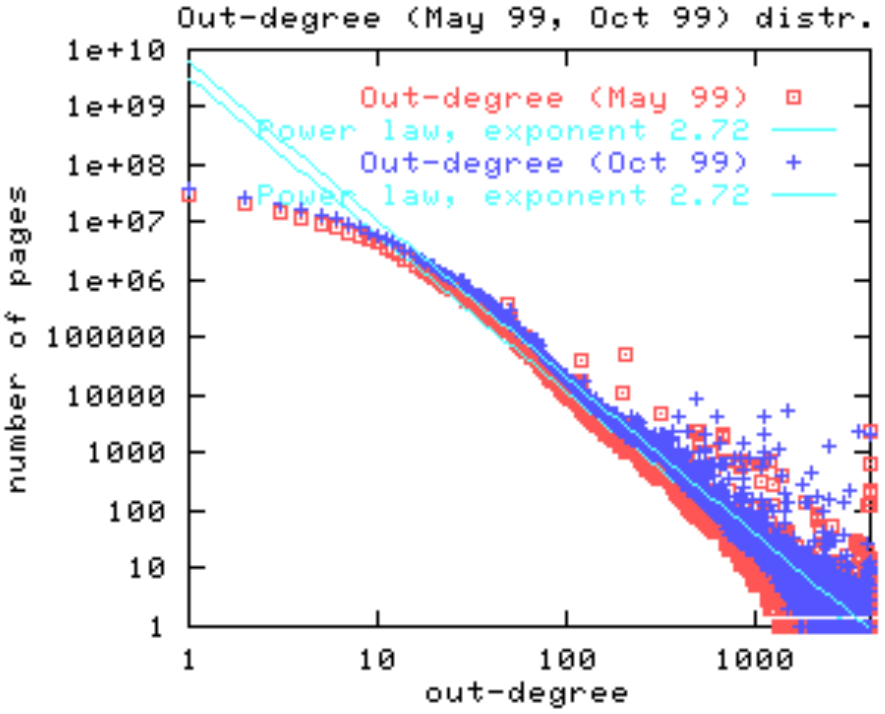


Internet domain topology

Node Degrees: Web

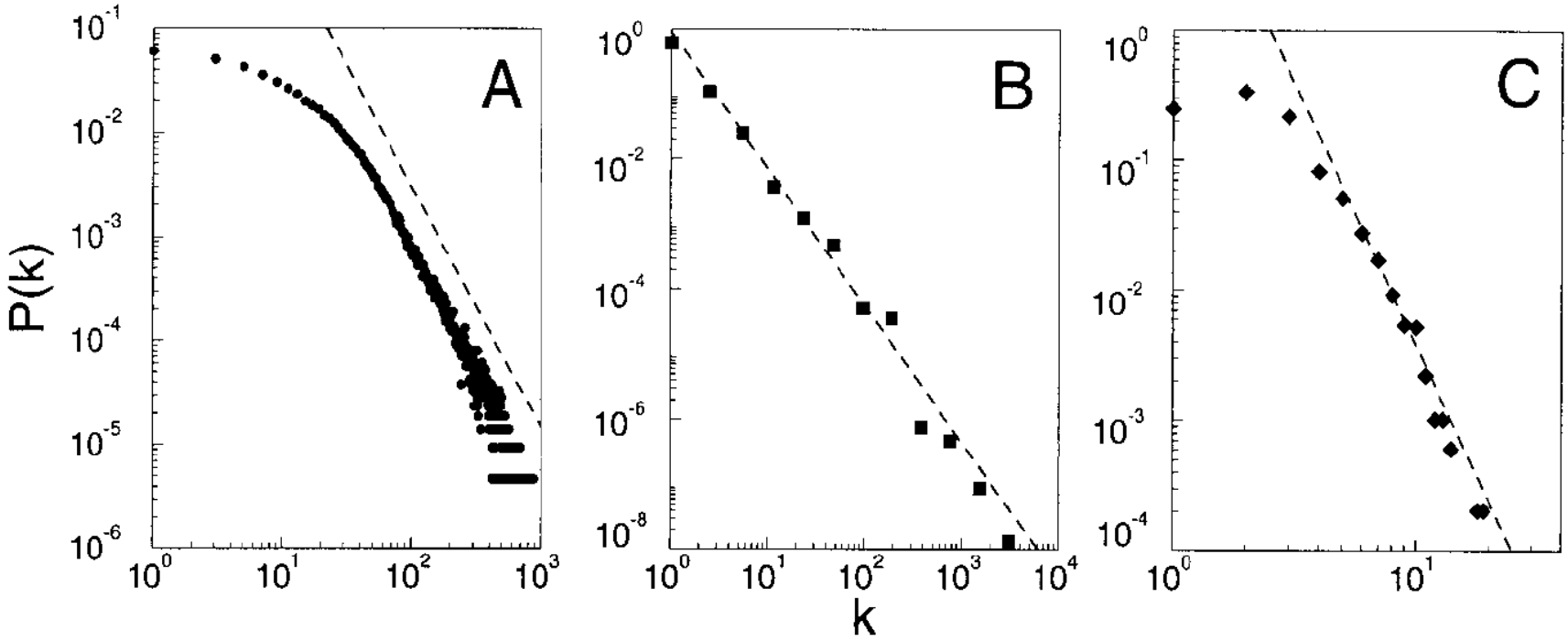
- **The World Wide Web** [Broder et al., 2000]



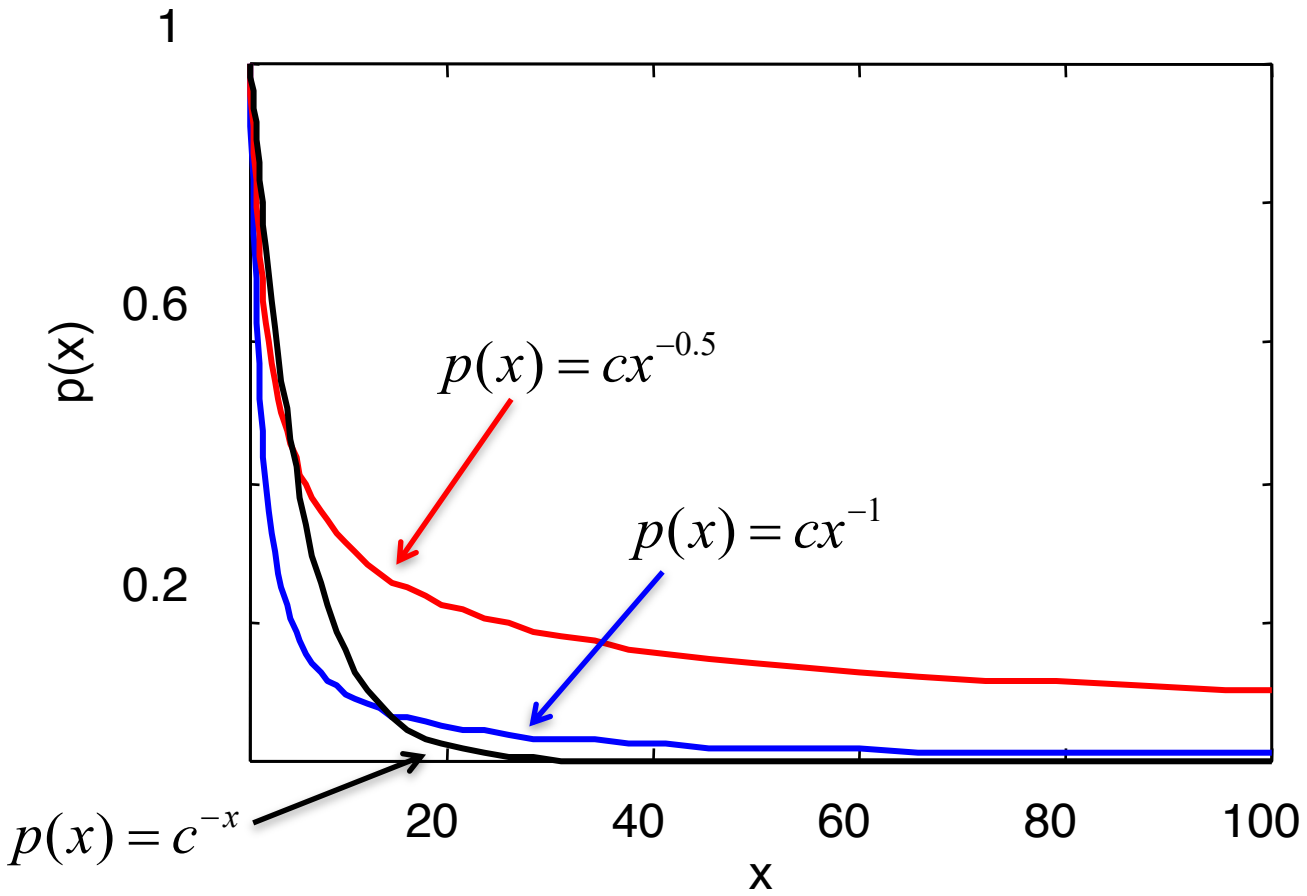


Node Degrees: Barabasi&Albert

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



Exponential vs. Power-Law

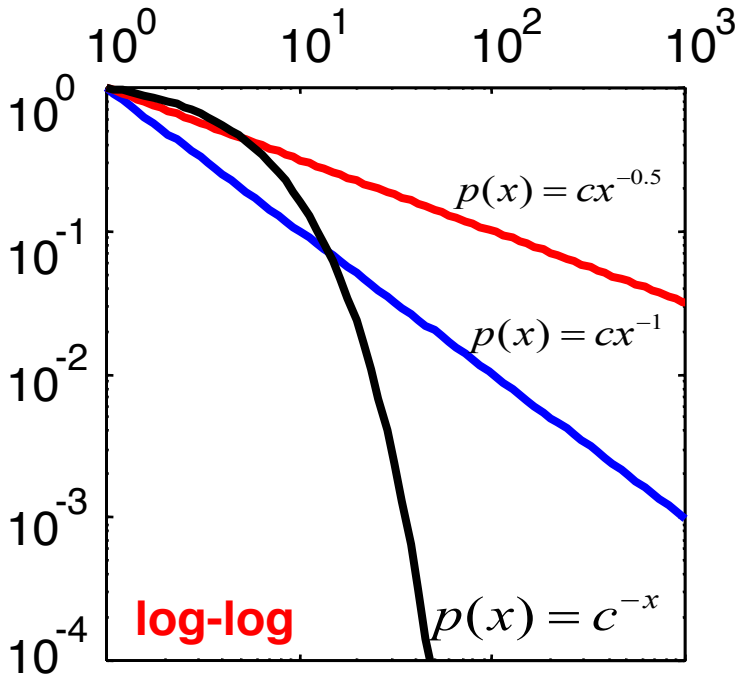


- Above a certain x value, the power law is always higher than the exponential!

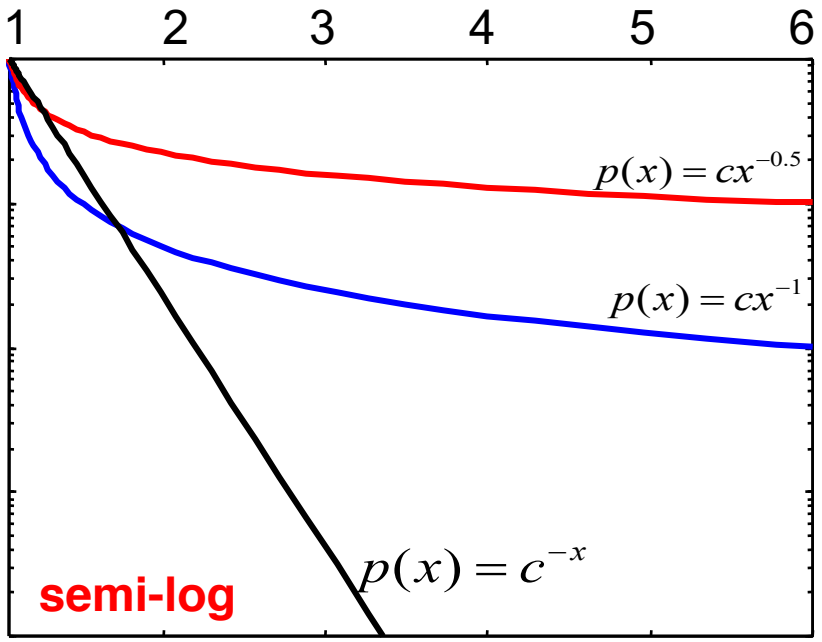
[Clauset-Shalizi-Newman 2007]

Exponential vs. Power-Law

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

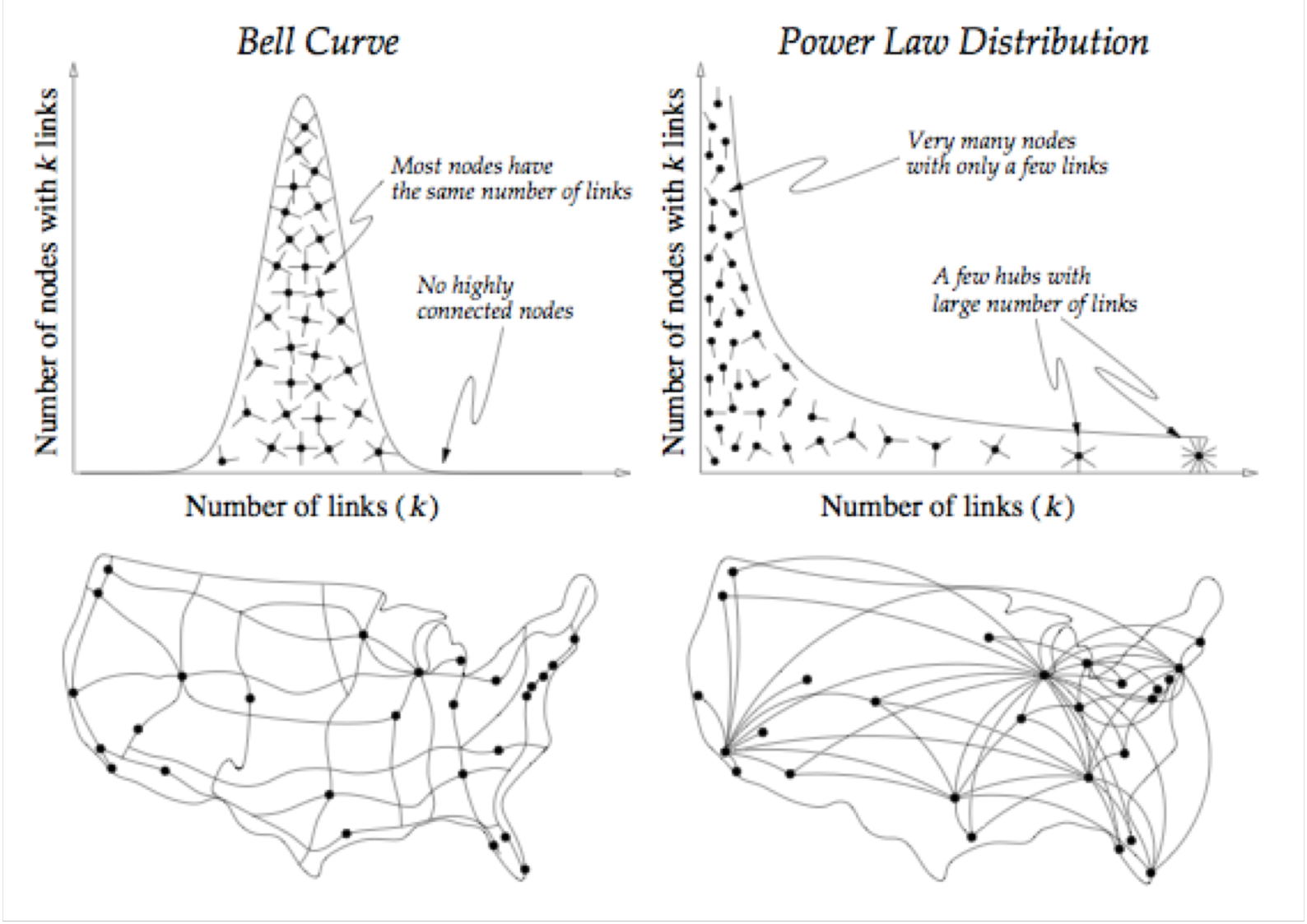


x … logarithmic axis
y … logarithmic axis



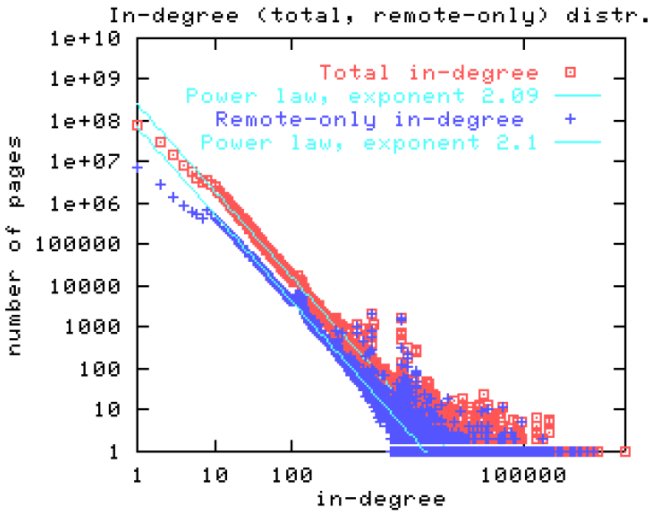
x … linear axis
y … logarithmic axis

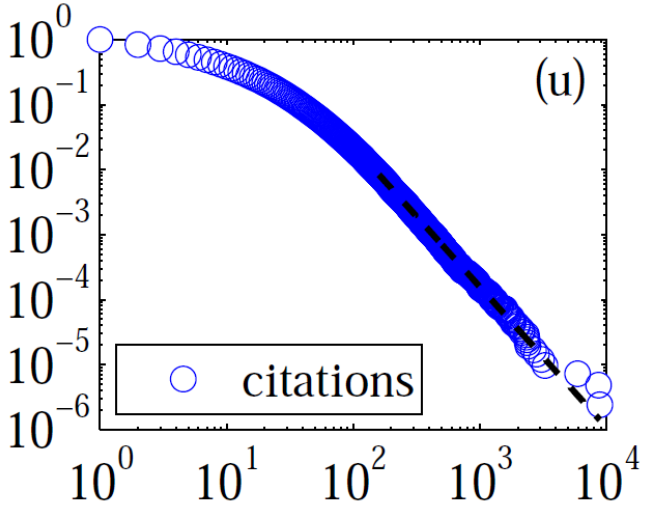
Exponential vs. Power-Law



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 scale-free come from?**
  - **Scale invariance:** There is no characteristic scale
    - **Scale invariance** means 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: $\boldsymbol{f(ax) = a^{\lambda} x^{\lambda} = a^{\lambda} f(x)}$

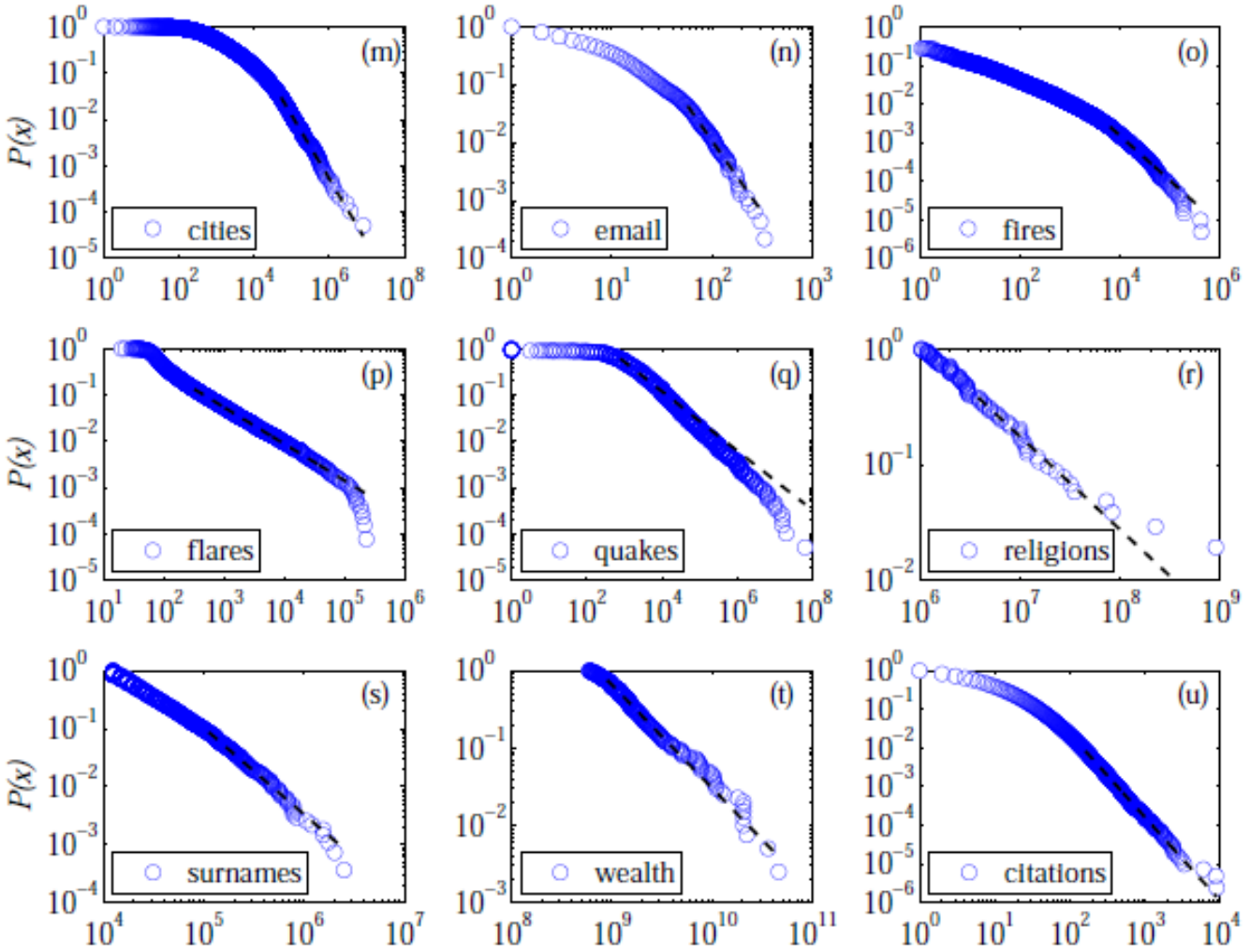
Log() or Exp() are not scale free!

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

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

[Clauset-Shalizi-Newman 2007]

Power-Laws are Everywhere

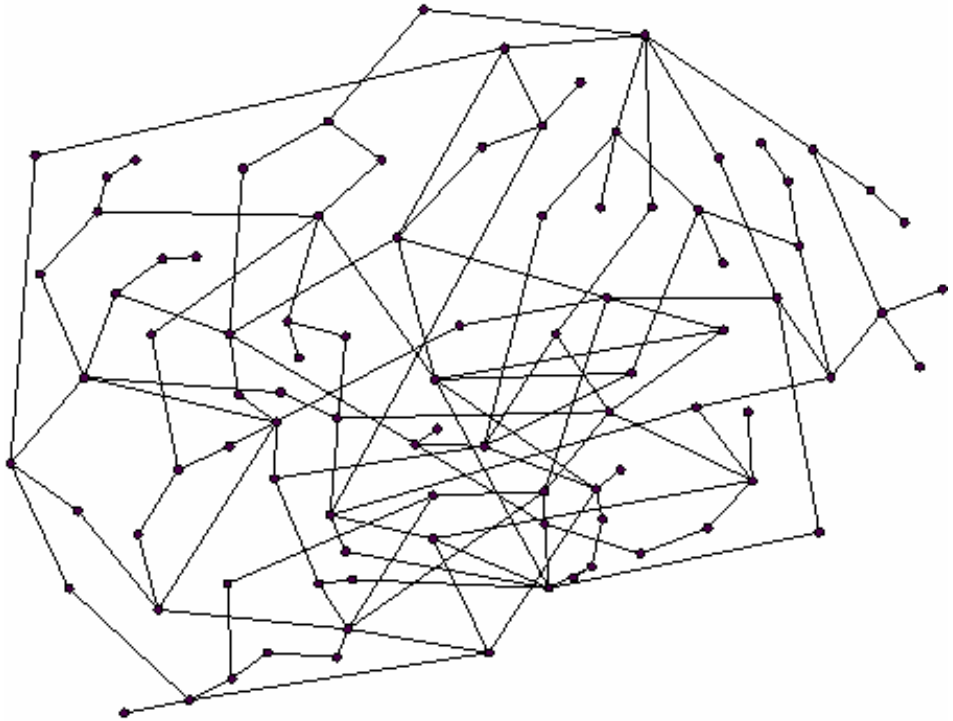


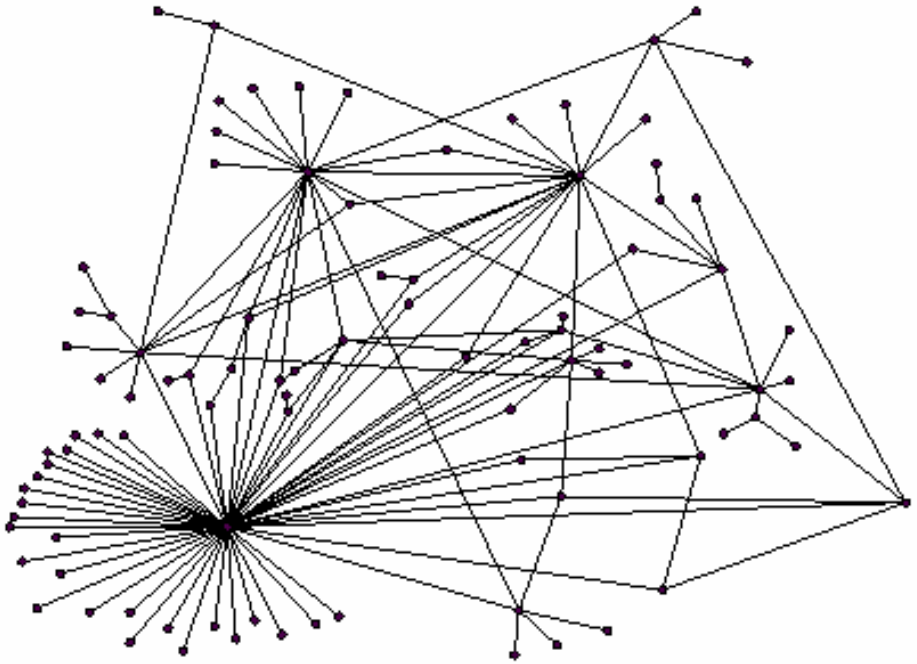
Many other quantities follow heavy-tailed distributions

Not Everyone Likes Power-Laws ☺

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

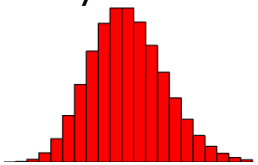
Random vs. Scale-free network





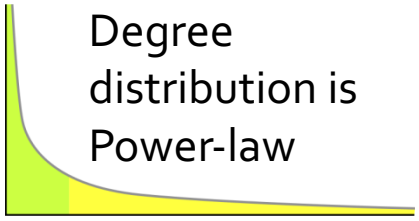
Random network

(Erdos-Renyi random graph)



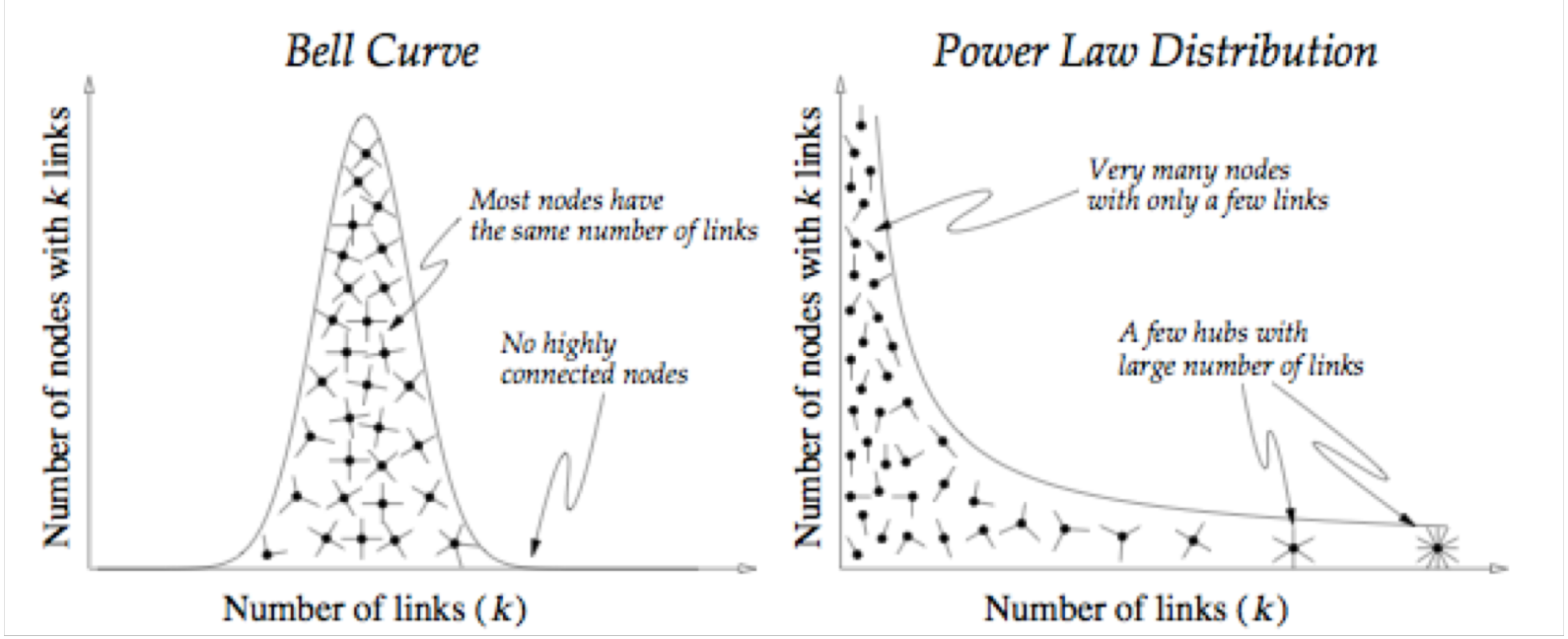
Degree distribution is Binomial

Scale-free (power-law) network



Preferential Attachment Model

Exponential vs. Power-Law Tails



Model: G_{np} **?**

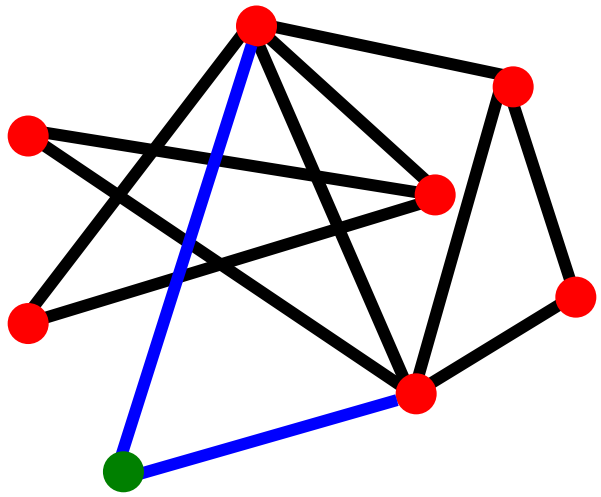
Rich Get Richer

- **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

Model: Preferential attachment

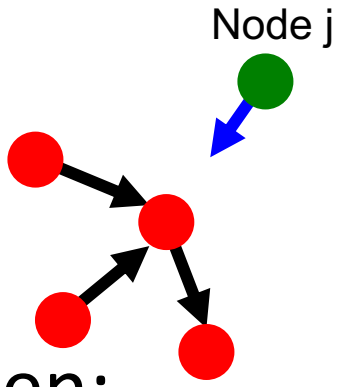
- **Preferential attachment:**
  [de Solla Price '65, Albert-Barabasi '99, Mitzenmacher '03]
  - Nodes arrive in order **1,2,…,n**
  - At step j, let d_i be the degree of node $i < j$
  - A new node j arrives and creates m out-links
  - Prob. of j linking to a previous node i is **proportional to degree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

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



The Exact Model

[Mitzenmacher, '03]



We analyze the following simple 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 Gives Power-Laws

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

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

Preferential attachment: Good news

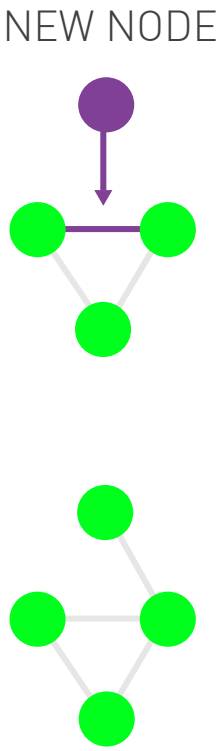
- **Preferential attachment gives power-law in-degrees!**
- Intuitively reasonable process
- Can tune model parameter p to get the observed exponent
  - On the web, $P[\text{node has in-degree } d] \sim d^{-2.1}$
  - $2.1 = 1+1/(1-p) \rightarrow p \sim 0.1$

Preferential Attachment: Bad News

- **Preferential attachment is not so good at predicting network structure**
  - **Age-degree correlation**
    - Node degree is proportional to its age
    - **Solution:** Node fitness (virtual degree)
  - **Links among high degree nodes:**
    - On the web nodes sometimes avoid linking to each other
- **Further questions:**
  - **What is a reasonable model for how people sample network nodes and link to them?**
    - Short random walks

Origins of Preferential Attachment

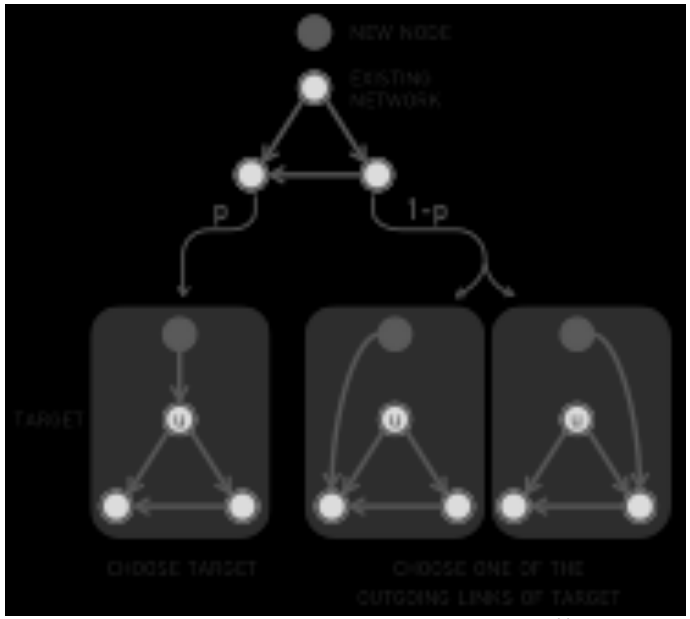
- **Link selection model** -- perhaps the simplest example of a local or random mechanism capable of generating preferential attachment
- **Growth:** At each time step we add a new node to the network
- **Link selection:** We select a link at random and connect the new node to one of nodes at the two ends of the selected link
- **This simple mechanism generates preferential attachment**
  - Why? Because node is picked with prob. proportional to the number of edges it has



Origins of Preferential Attachment

Copying model:

- **(a) Random Connection:** with prob. p the new node links to random u
- **(b) Copying:** With prob. $1 - p$ randomly choose an outgoing link of node u and connect the new node to the selected link's target
  - The new node "copies" one of the links of an earlier node



Origins of Preferential Attachment

Analysis of the copying model:

- **(a)** the probability of selecting a node is 1/N
- **(b)** is equivalent to selecting a node linked to a randomly selected link. The probability of selecting a degree-k node through the copying process of step (b) is $k/2E$ for undirected networks
- Again, the likelihood that the new node will connect to a degree-k node follows preferential attachment

Examples:

- **Social networks:** Copy your friend's friends.
- **Citation Networks:** Copy references from papers we read
- **Protein interaction networks:** gene duplication

Many models lead to Power-Laws

- **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 and attach to both endpoints of this edge
- **Node duplication**
  - Duplicate a node with all its edges
  - Randomly prune edges of new node

Distances in Preferential Attachment

$$\bar{h} = \begin{cases} const & \alpha = 2 \\ \frac{\log\log n}{\log(\alpha - 1)} & 2 < \alpha < 3 \\ \frac{\log n}{\log\log n} & \alpha = 3 \\ \log n & \alpha > 3 \end{cases}$$

Ultra small world

Small world

Avg. path length

Degree exponent

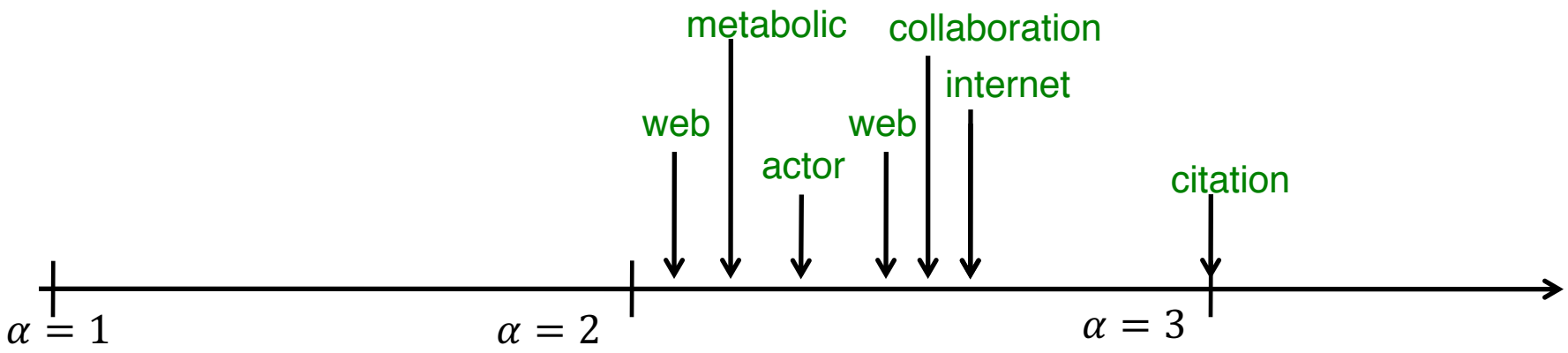
Size of the biggest hub is of order *O(N)*. Most nodes can be connected within two steps, thus the average path length will be independent of the network size *n*.

The avg. path length increases slower than logarithmically with *n*. In G_{np} all nodes have comparable degree, thus most paths will have comparable length. In a scale-free network vast majority of the paths go through the few high degree hubs, reducing the distances between nodes.

Some models produce $\alpha = 3$. This was first derived by Bollobas et al. for the network diameter in the context of a dynamical model, but it holds for the average path length as well.

The second moment of the distribution is finite, thus in many ways the network behaves as a random network. Hence the average path length follows the result that we derived for the random network model earlier.

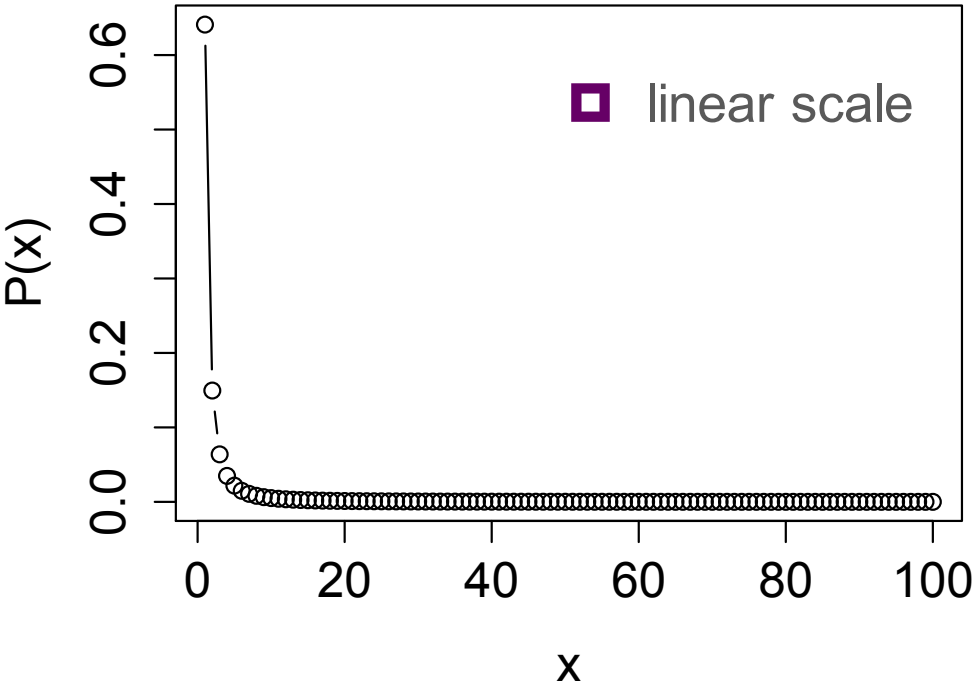
Summary: Scale-Free Networks

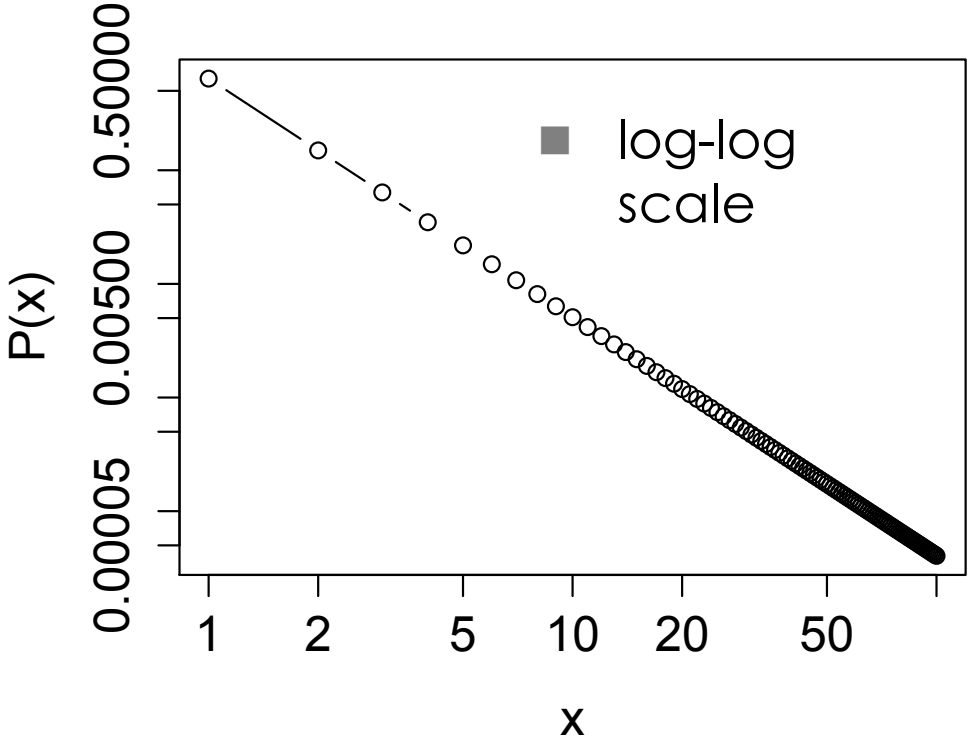


Second moment $\langle k^2 \rangle$ diverges	$\langle k^2 \rangle$ finite
Average $\langle k \rangle$ diverges	Average $\langle k \rangle$ finite
Ultra small world behavior	Small world

Regime full of anomalies…	The scale-free behavior is relevant	Behaves like a random network

Power-law distribution





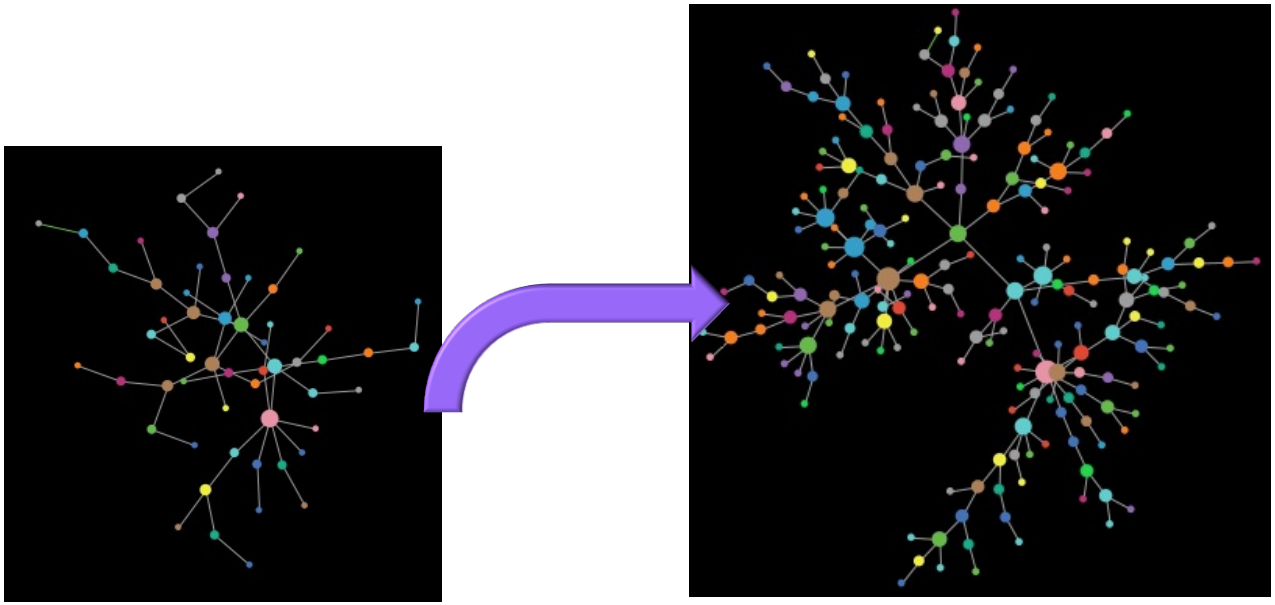
- high skew (asymmetry)
- straight line on a log-log plot

Quiz Q:

- As the exponent α increases, the downward slope of the line on a log-log plot
  - stays the same
  - becomes milder
  - becomes steeper

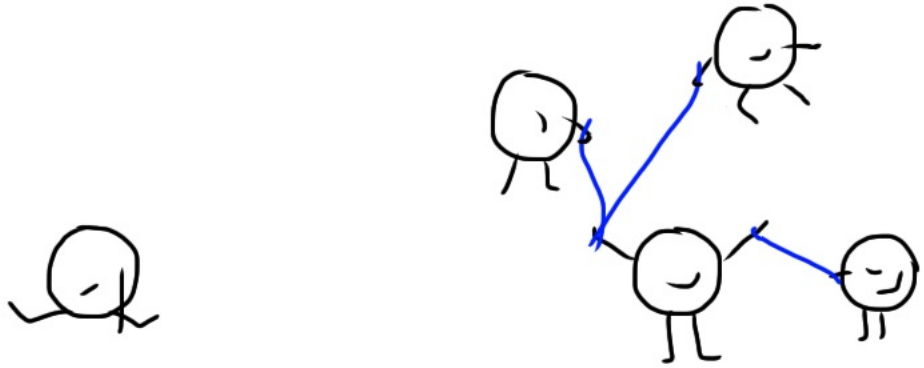
2 ingredients in generating power-law networks

- nodes appear over time (growth)

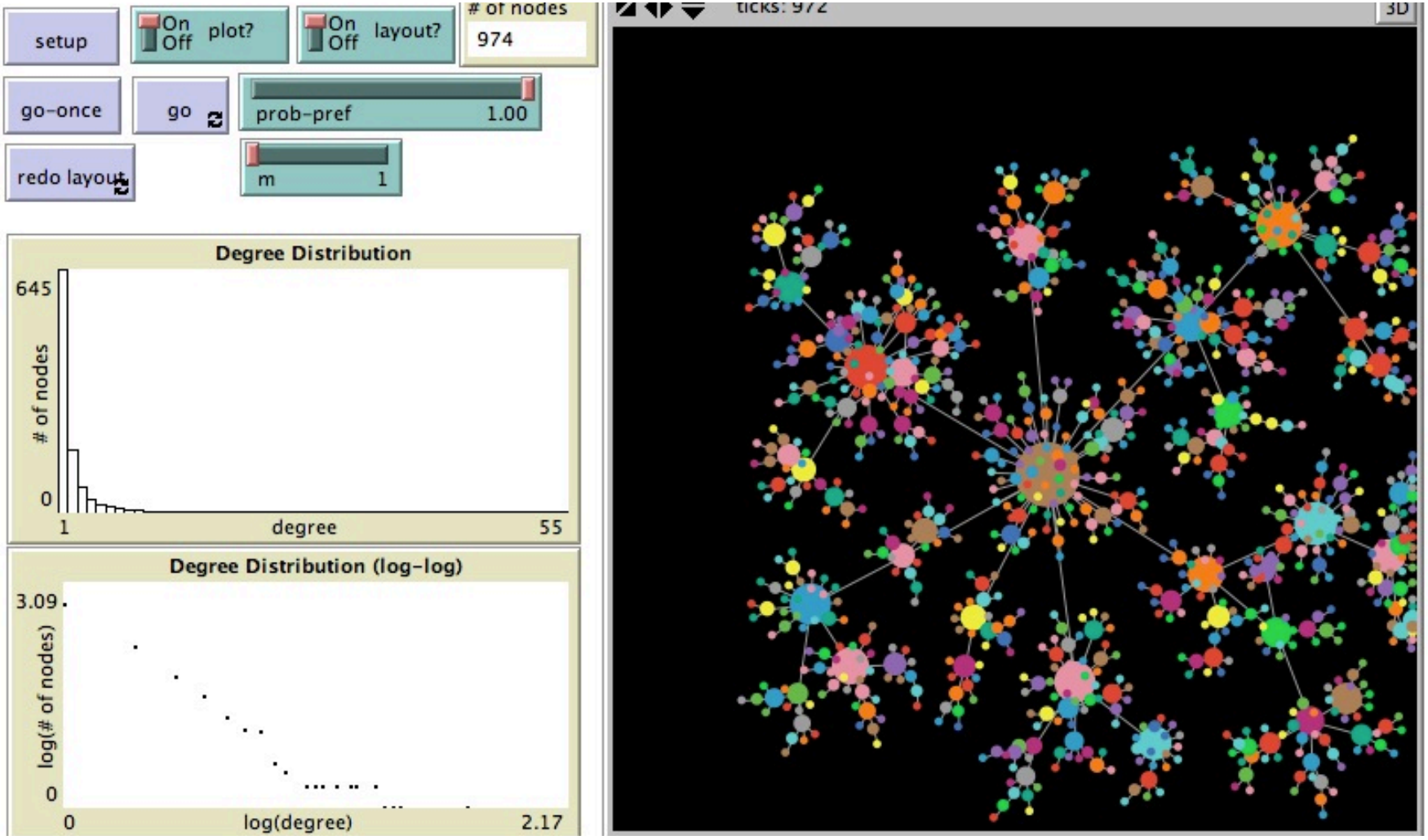


2 ingredients in generating power-law networks

- nodes prefer to attach to nodes with many connections (preferential attachment, cumulative advantage)



try it yourself



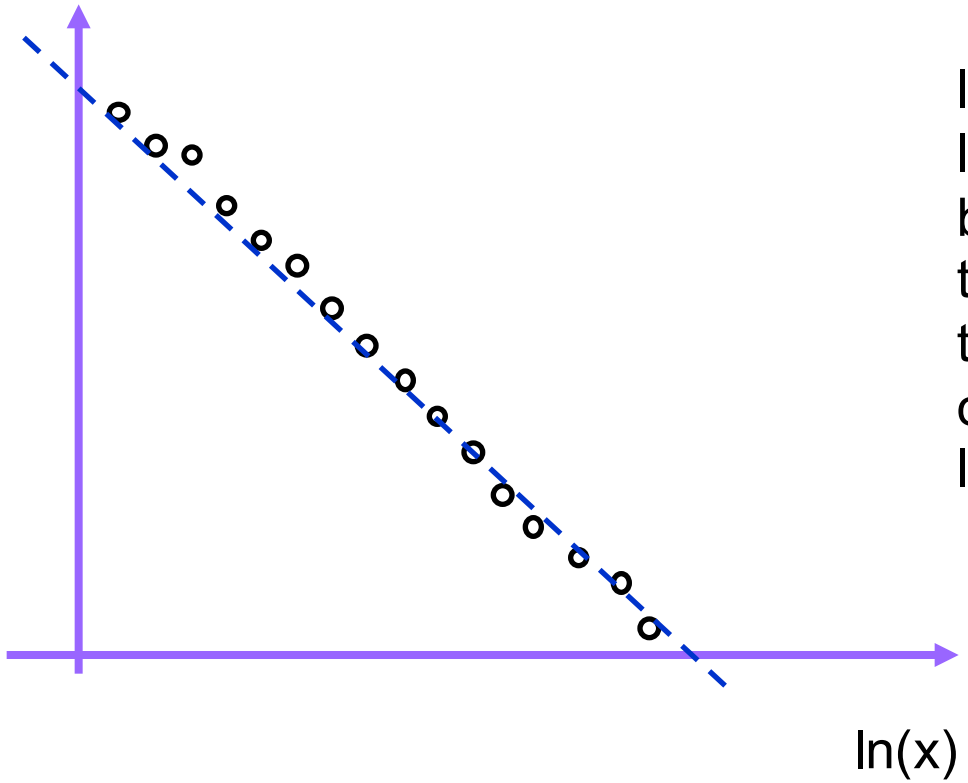
http://web.stanford.edu/class/cs224w/NetLogo/RAndPrefAttachment.nlogo

Quiz Q:

- Relative to the random growth model, the degree distribution in the preferential attachment model
  - resembles a power-law distribution less
  - resembles a power-law distribution more

Fitting power-law distributions

- Most common and not very accurate method:
  - Bin the different values of x and create a frequency histogram



ln(x) is the natural logarithm of x, but any other base of the logarithm will give the same exponent of α because $\log_{10}(x) = \ln(x)/\ln(10)$

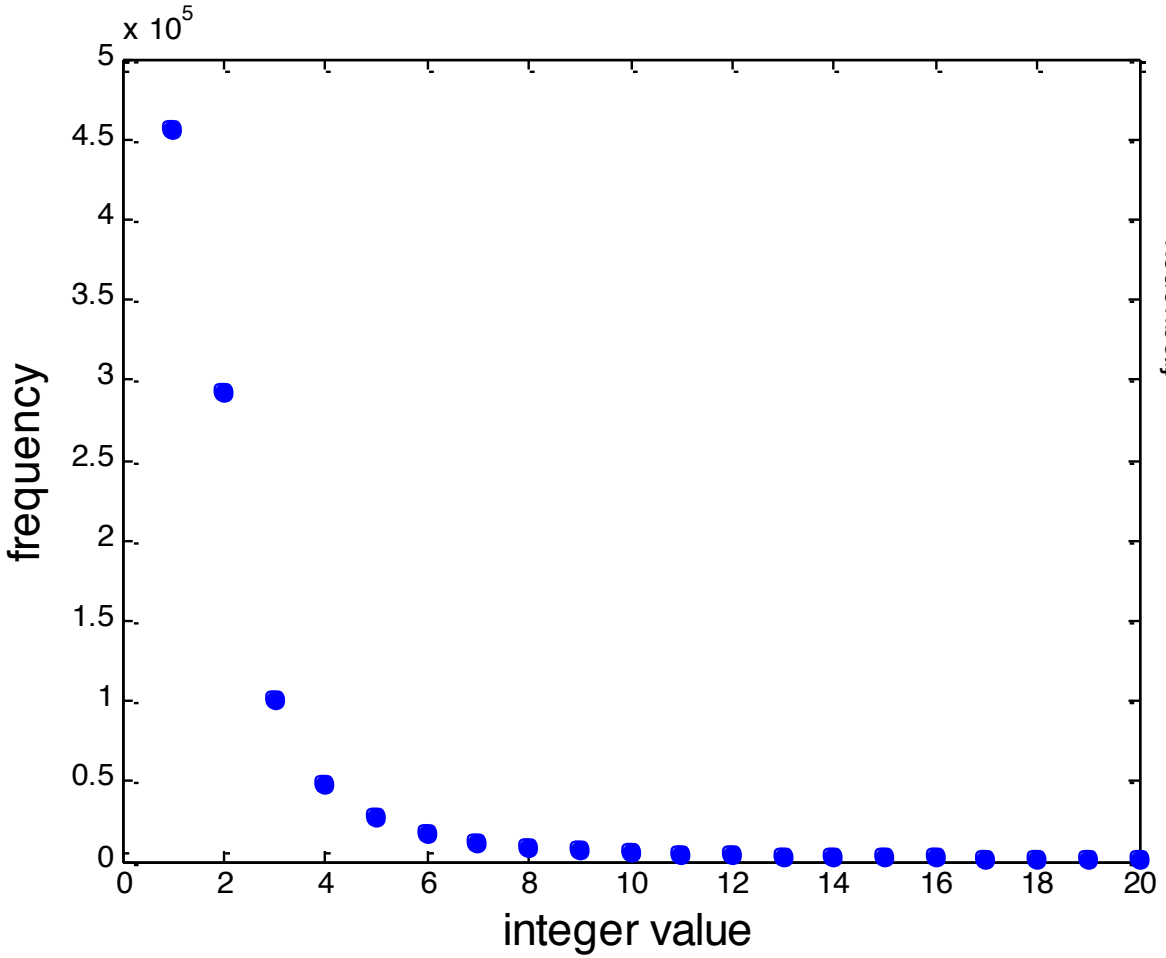
x can represent various quantities, the indegree of a node, the magnitude of an earthquake, the frequency of a word in text

Example on an artificially generated data set

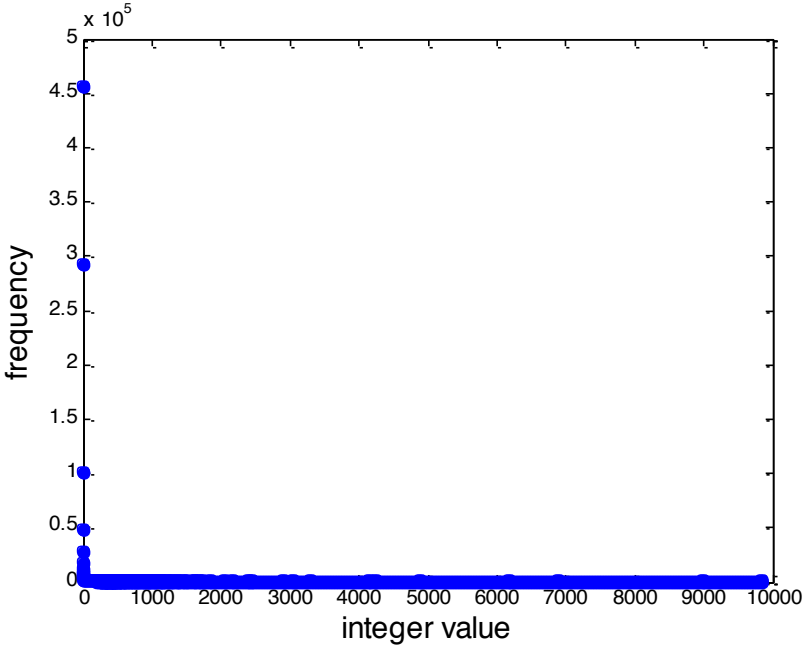
- Take 1 million random numbers from a distribution with $\alpha = 2.5$
- Can be generated using the so-called 'transformation method'
- Generate random numbers r on the unit interval $0 \le r < 1$
- then $x = (1-r)^{-1/(\alpha-1)}$ is a random power law distributed real number in the range $1 \le x < \infty$

Linear scale plot of straight bin of the data

- Number of times 1 or 3843 or 99723 occurred
- Power-law relationship not as apparent
- Only makes sense to look at smallest bins



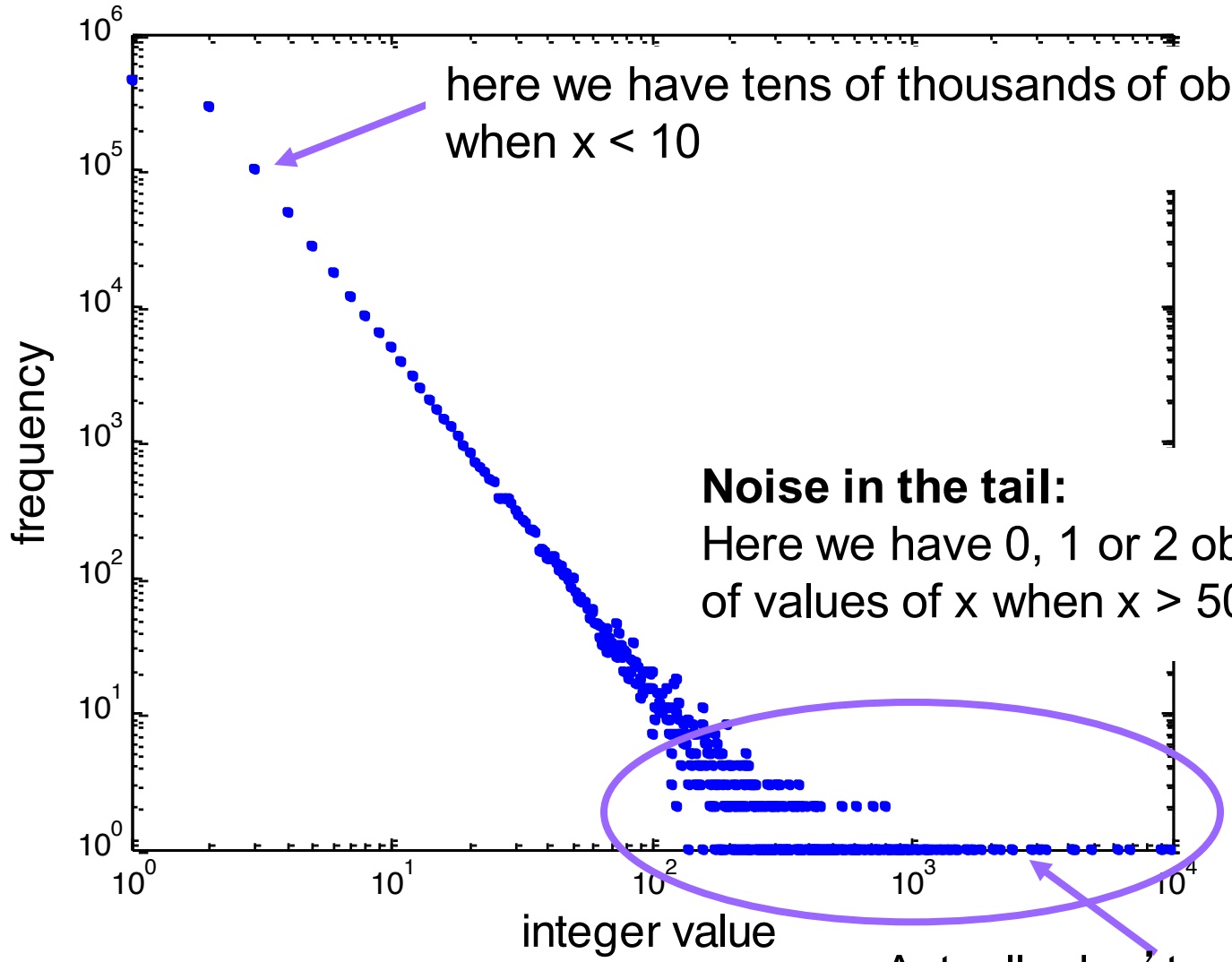
first few bins



whole range

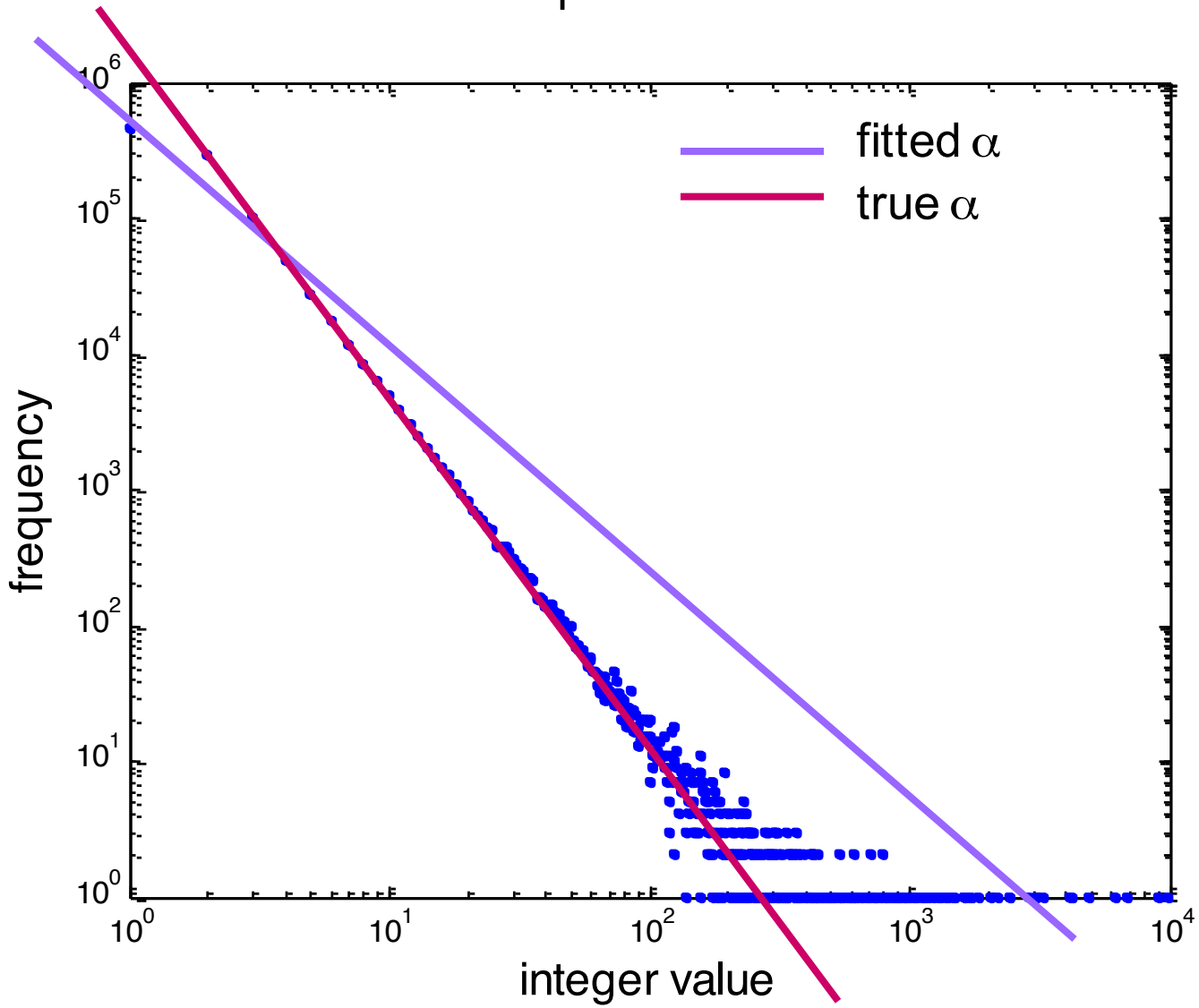
Log-log scale plot of simple binning of the data

- Same bins, but plotted on a log-log scale



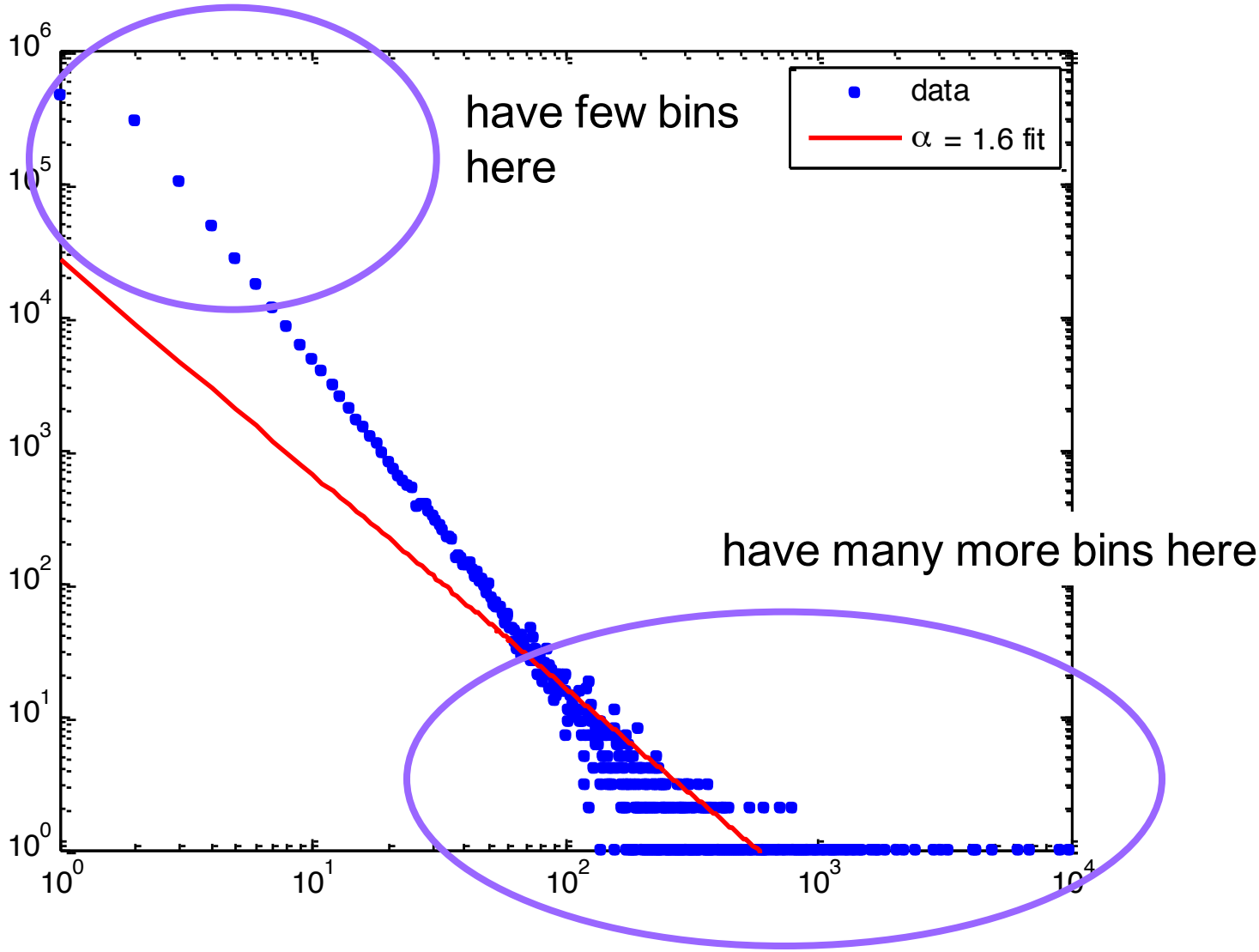
Log-log scale plot of straight binning of the data

- Fitting a straight line to it via least squares regression will give values of the exponent α that are too low



What goes wrong with straightforward binning

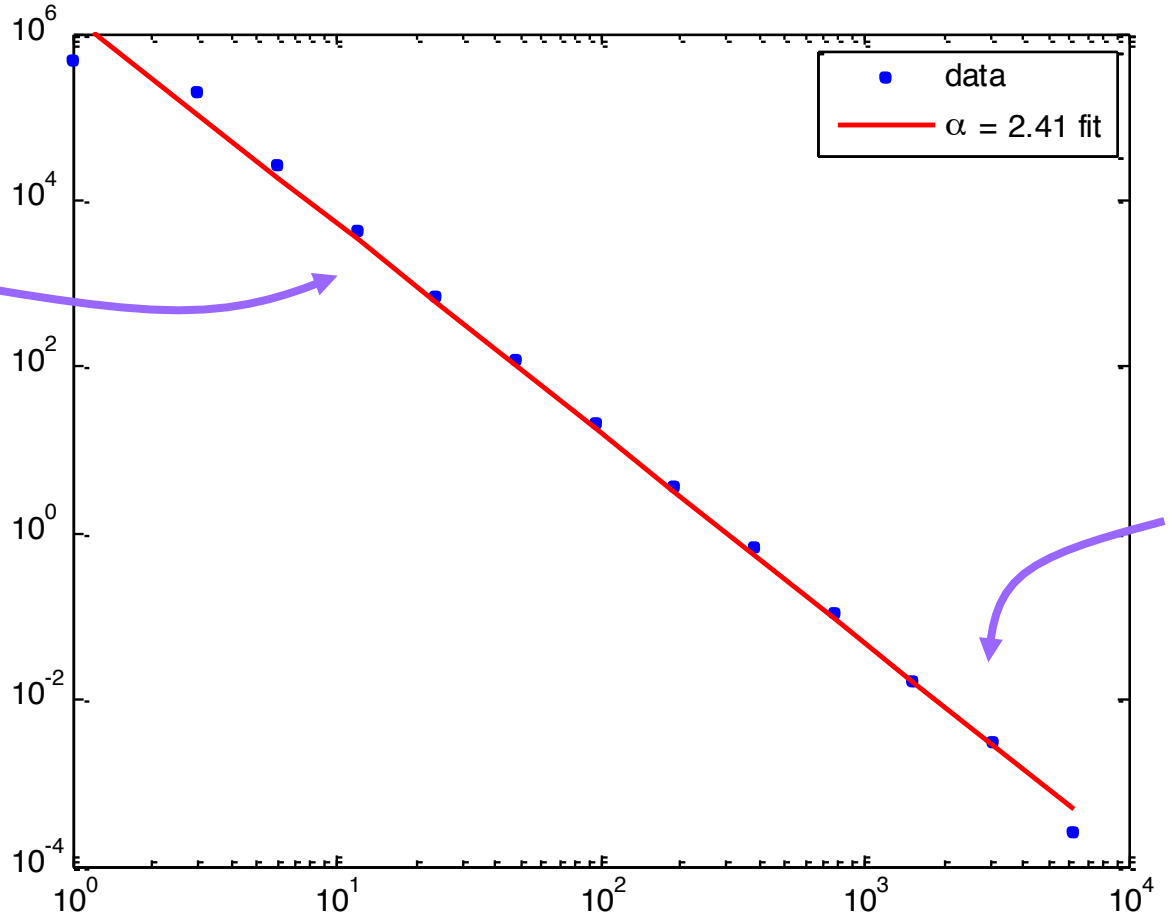
- Noise in the tail skews the regression result



First solution: logarithmic binning

- bin data into exponentially wider bins:
  - 1, 2, 4, 8, 16, 32, …
- normalize by the width of the bin

evenly spaced datapoints



less noise in the tail of the distribution

- disadvantage: binning smoothes out data but also loses information

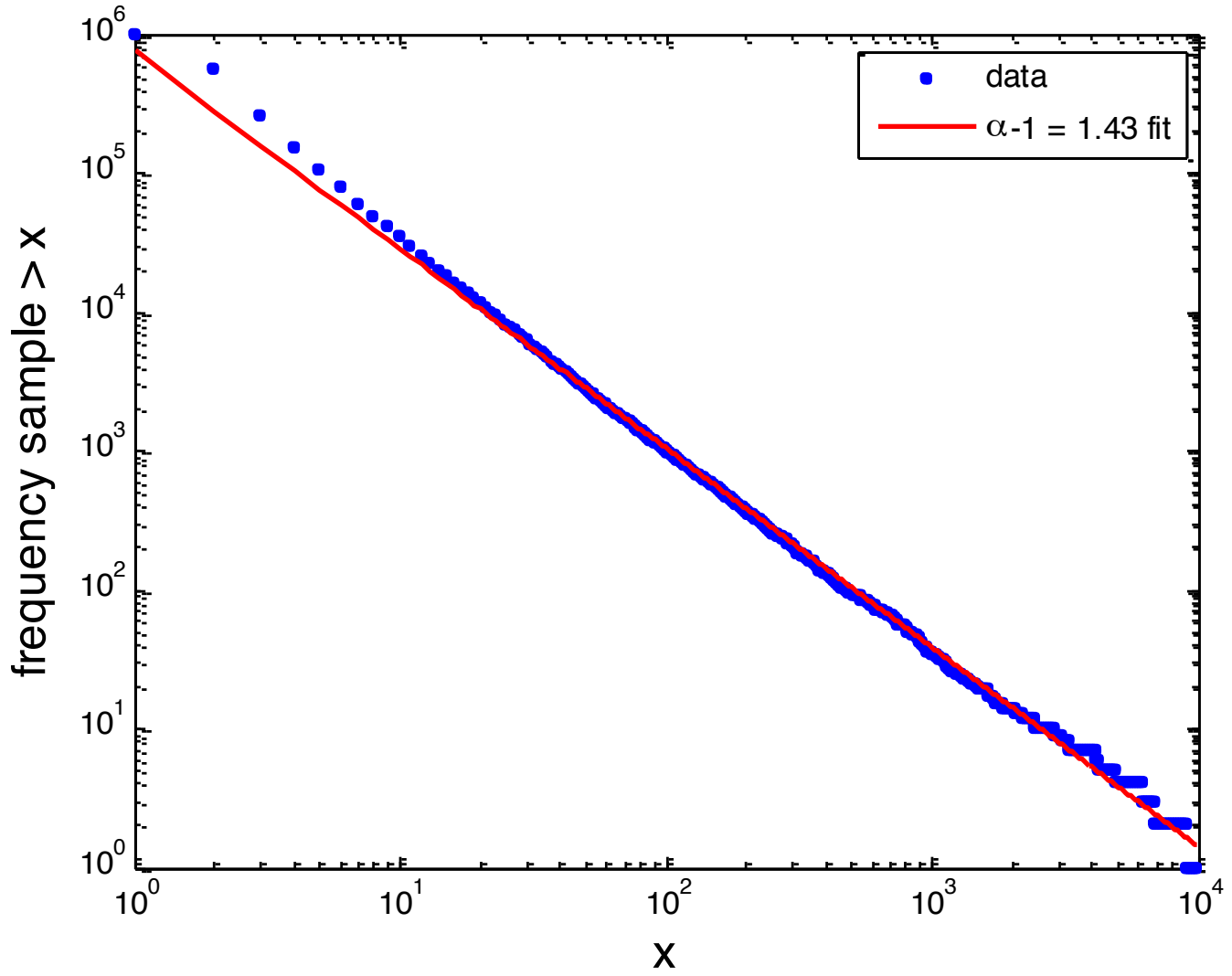
Second solution: cumulative binning

- No loss of information
  - No need to bin, has value at each observed value of x
- But now have cumulative distribution
  - i.e. how many of the values of x are at least X
  - The cumulative probability of a power law probability distribution is also power law but with an exponent **α - 1**

$$\int cx^{-\alpha} = \frac{c}{1-\alpha} x^{-(\alpha-1)}$$

Fitting via regression to the cumulative distribution

- fitted exponent (2.43) much closer to actual (2.5)

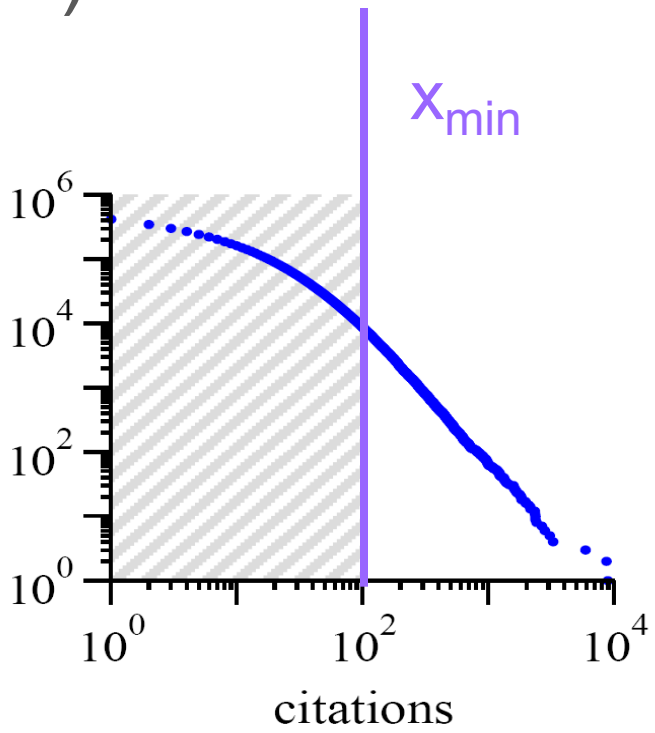


Where to start fitting?

- some data exhibit a power law only in the tail
- after binning or taking the cumulative distribution you can fit to the tail
- so need to select an x_{min} the value of x where you think the power-law starts
- certainly x_{min} needs to be greater than 0, because $x^{-\alpha}$ is infinite at x = 0

Example:

- Distribution of citations to papers
- power law is evident only in the tail (x_{min} > 100 citations)



Source:MEJ Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

Maximum likelihood fitting – best

- You have to be sure you have a power-law distribution (this will just give you an exponent but not a goodness of fit)

$$\alpha = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1}$$

- x_i are all your data points, and you have *n* of them
- for our data set we get $\alpha = 2.503$ – pretty close!

Some exponents for real world data

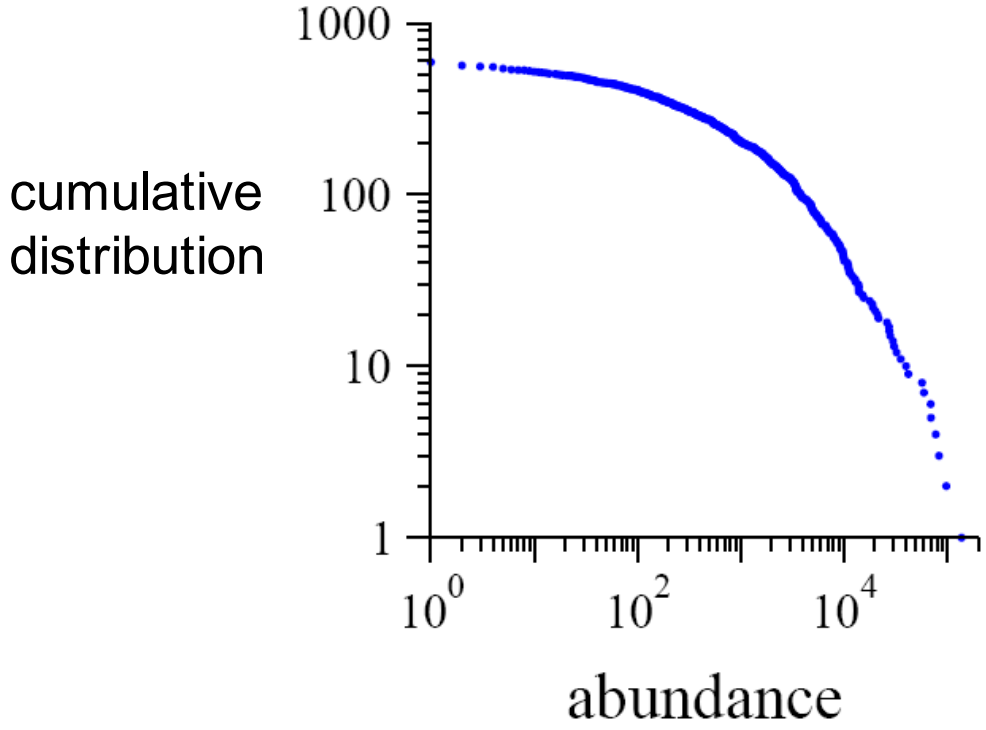
	x_{min}	exponent α
frequency of use of words	1	2.20
number of citations to papers	100	3.04
number of hits on web sites	1	2.40
copies of books sold in the US	2 000 000	3.51
telephone calls received	10	2.22
magnitude of earthquakes	3.8	3.04
diameter of moon craters	0.01	3.14
intensity of solar flares	200	1.83
intensity of wars	3	1.80
net worth of Americans	$600m	2.09
frequency of family names	10 000	1.94
population of US cities	40 000	2.30

Many real world networks are power law

	exponent α (in/out degree)
film actors	2.3
telephone call graph	2.1
email networks	1.5/2.0
sexual contacts	3.2
WWW	2.3/2.7
internet	2.5
peer-to-peer	2.1
metabolic network	2.2
protein interactions	2.4

Hey, not everything is a power law

- number of sightings of 591 bird species in the North American Bird survey in 2003.



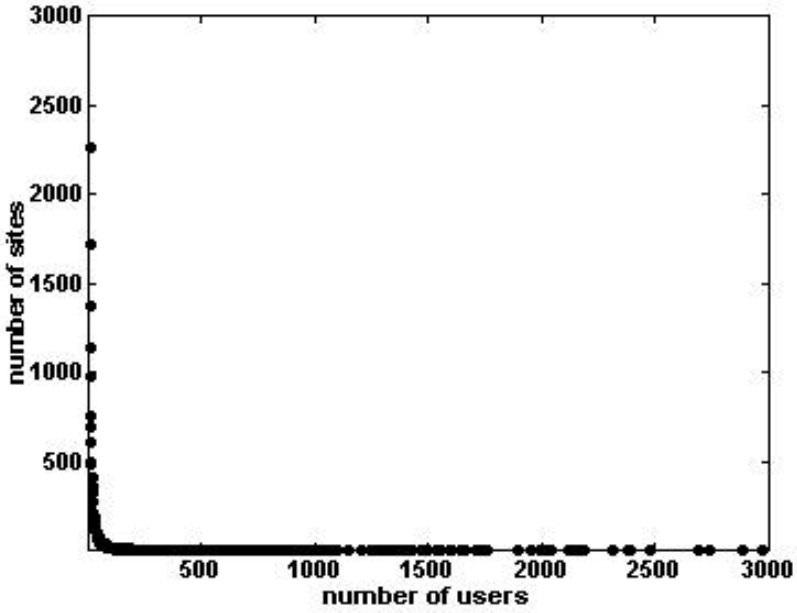
- another example:
  - size of wildfires (in acres)

Source:MEJ Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

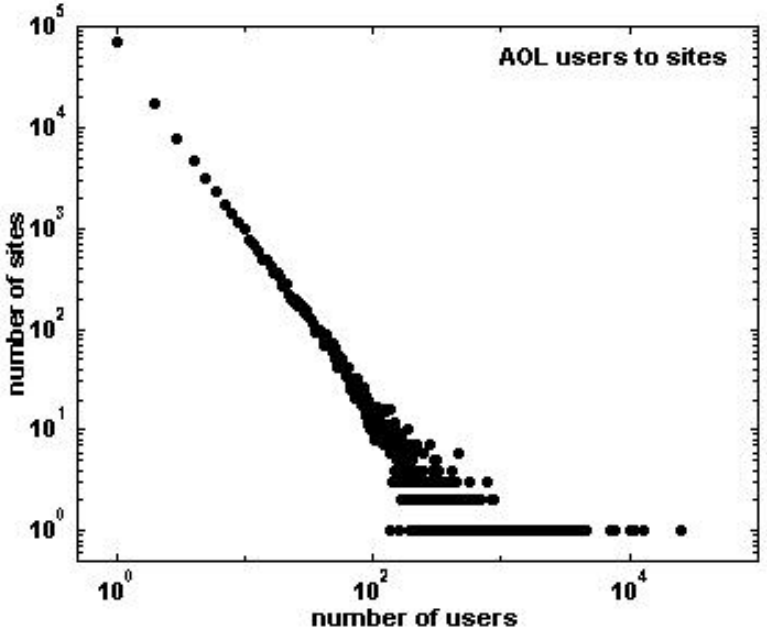
Not every network is power law distributed

- reciprocal, frequent email communication
- power grid
- Roget's thesaurus
- company directors…

Example on a real data set: number of AOL visitors to different websites back in 1997



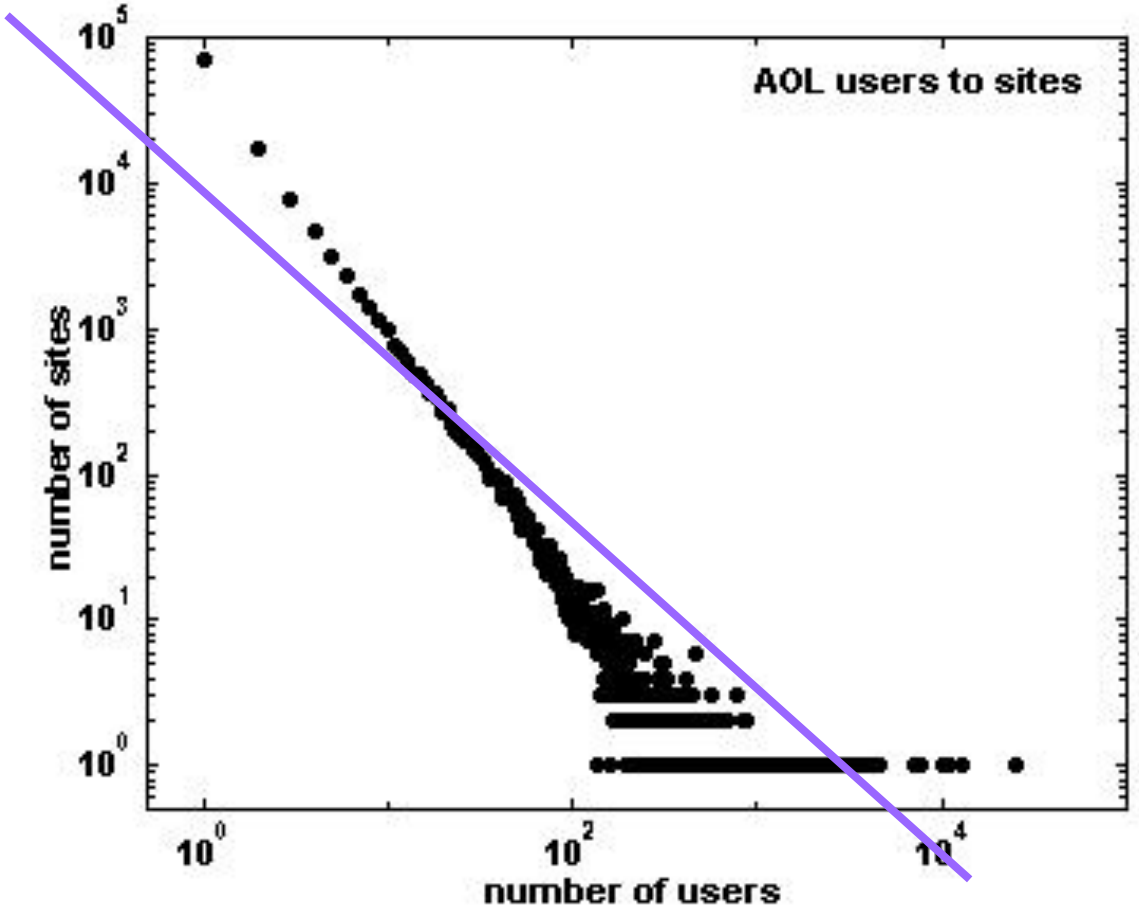
simple binning on a linear scale



simple binning on a log-log scale

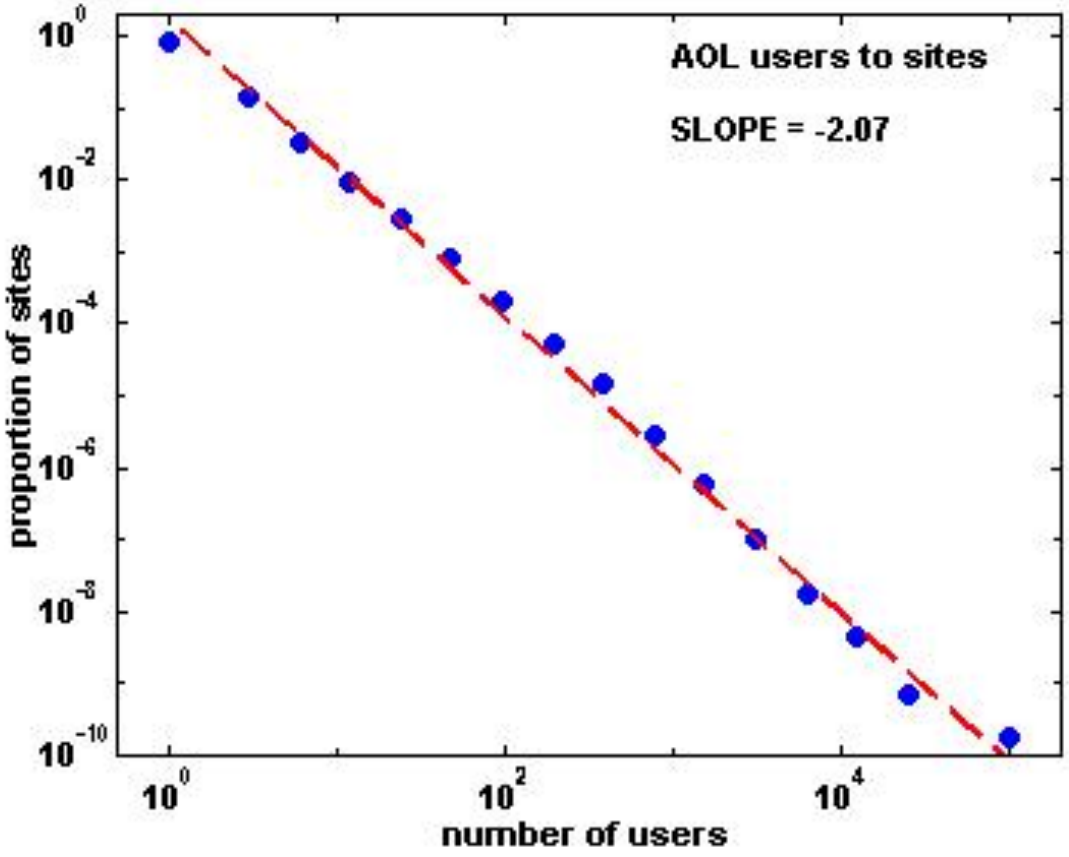
trying to fit directly…

- direct fit is too shallow: $\alpha = 1.17$…



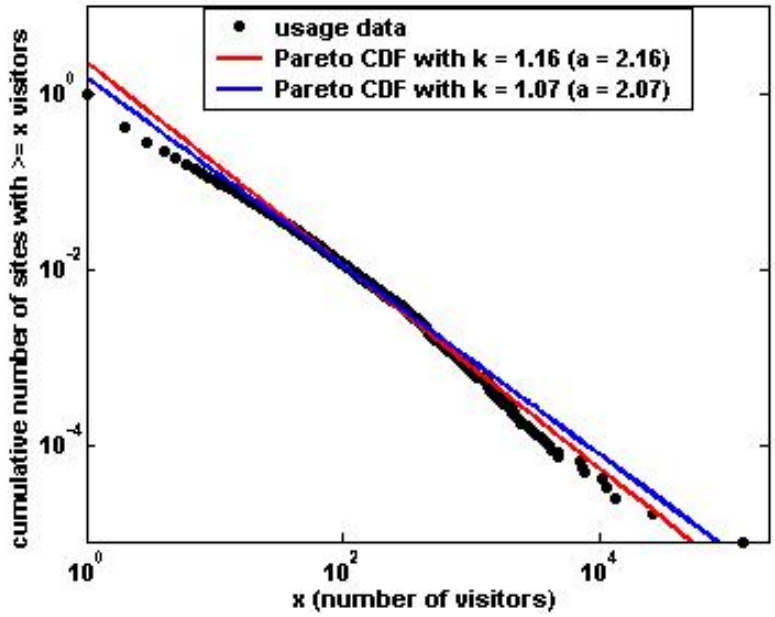
Binning the data logarithmically helps

- select exponentially wider bins
  - 1, 2, 4, 8, 16, 32, ….



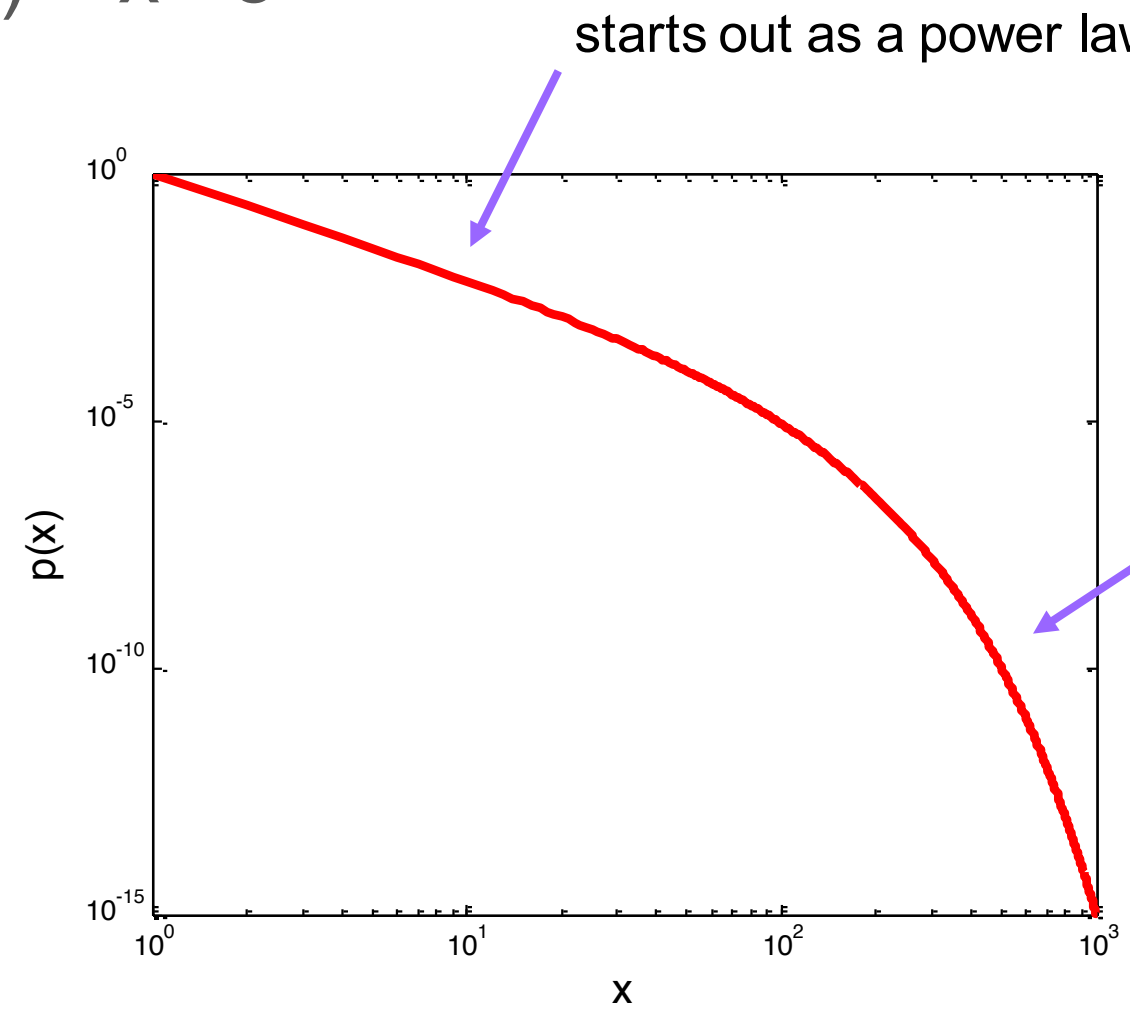
Or we can try fitting the cumulative distribution

- Shows perhaps 2 separate power-law regimes that were obscured by the exponential binning
- Power-law tail may be closer to 2.4



Another common distribution: power-law with an exponential cutoff

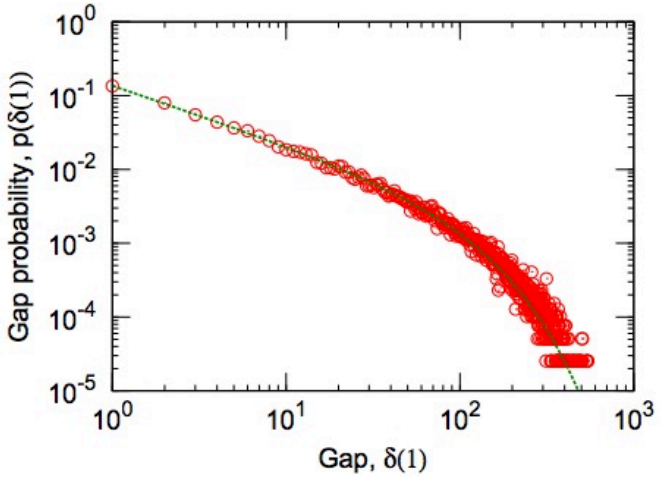
- $p(x) \sim x^{-a} e^{-x/\kappa}$



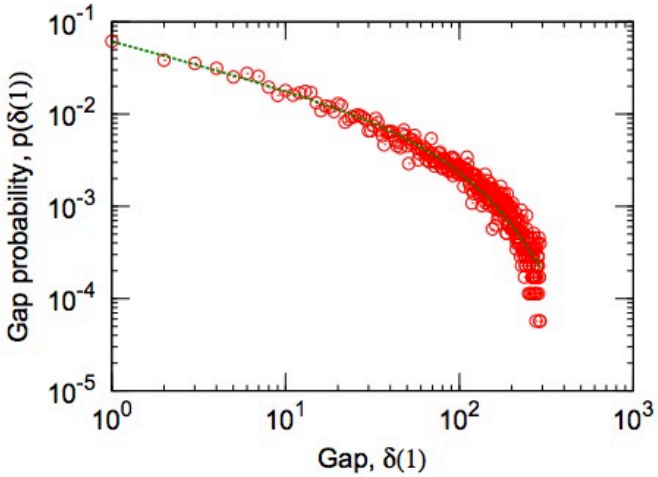
but could also be a lognormal or double exponential…

example: time between edge initiations

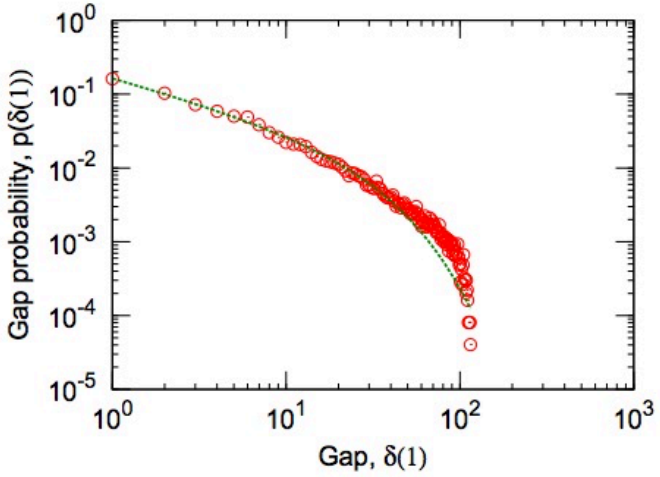
Q: Why is the cutoff present?



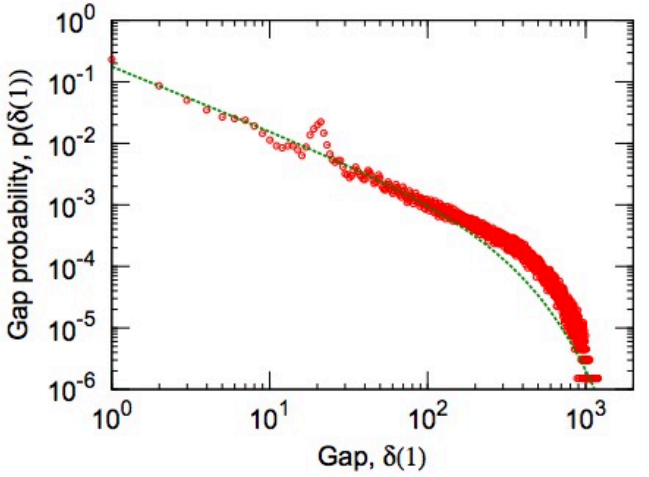
(a) FLICKR



(b) DELICIOUS



(c) ANSWERS



(d) LINKEDIN

Leskovec et al., KDD'08

Wrap up on power-laws

- Power-laws are cool and intriguing
- But make sure your data is actually power-law before boasting