CentraleSupélec

ESSEC BUSINESS SCHOOL

Network Science Analytics

Option Applied Math and M.Sc. in DSBA

Lecture 2A

Random graphs and the small-world phenomenon

Fragkiskos Malliaros

Friday, January 26, 2018

Acknowledgements

- The lecture is partially based on material by
  - Jure Leskovec, Stanford University
  - Aaron Clauset, CU Boulder
  - Manos Papaggelis, 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!

Last time, we studied 3 basic properties of the MSN messenger network

Key Network Properties

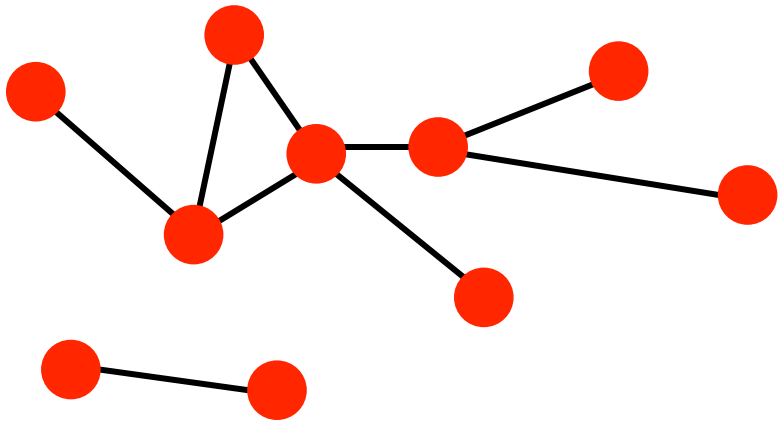
Degree distribution:	$P(k)$
Path length:	h
Clustering coefficient:	C

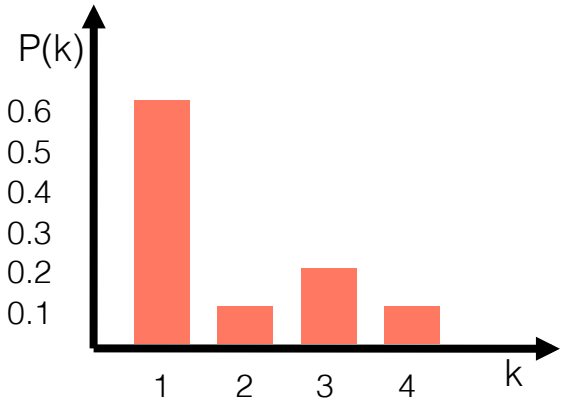
Degree Distribution

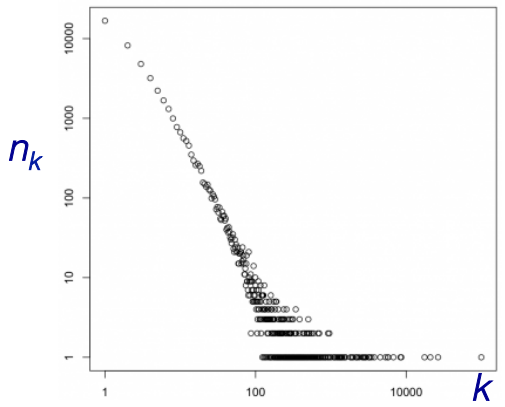
- Degree distribution *P(k)*: Probability that a randomly chosen node has degree *k*

  n_k = # nodes with degree *k*

- Normalized histogram:
  $P(k) = n_k / n$ ➔ plot







Number of Paths

Property: Let A^h denote the h-th power of A, with entries A^h_{uv}

- Then, element A^h_{uv} captures the number of $u - v$ paths of length h in G

  - # of paths of length $h=1$: If there is a link between u and v, $A_{uv}=1$ else $A_{uv}=0$

  - # of paths of length $h=2$: If there is a path of length two between u and v through k, the product $A_{uk} A_{kv}=1$ else $A_{uk} A_{kv}=0$

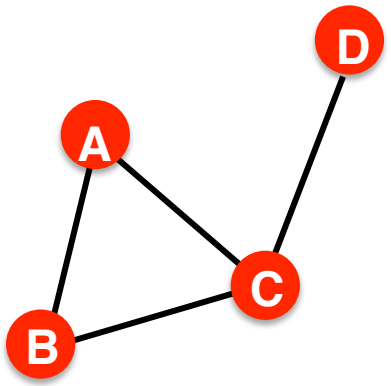
$$H^{(2)}_{uv} = \sum_{k=1}^{n} \mathbf{A}_{uk}\mathbf{A}_{kv} = \mathbf{A}^2_{uv}$$

  - # of paths of length h: If there is a path of length h between u and v then $A_{uk} .... A_{kv}=1$ else $A_{uk} .... A_{kv}=0$

$$H^{(h)}_{uv} = \mathbf{A}^h_{uv}$$

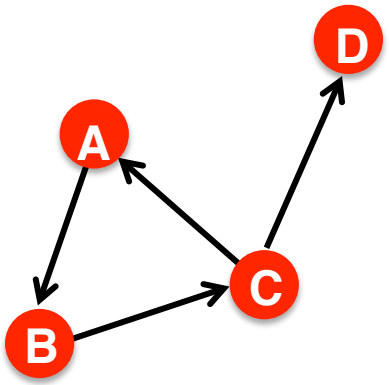
(holds for both directed and undirected graphs)

Distance in a Graph



$h_{B,D} = 2$

- 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 disconnected, the distance is usually defined as infinite



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

- In directed graphs paths need to follow the direction of the arrows
  - Consequence: Distance is **not symmetric**: $h_{A,C} \neq h_{C,A}$

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

$$\bar{h} = \frac{1}{n(n-1)} \sum_{i,j \neq i} h_{ij}$$

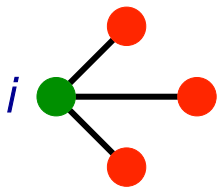
where h_{ij} is the distance from node i to node j

Clustering Coefficient (1/2)

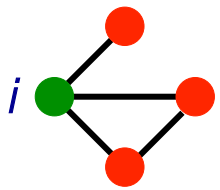
- Clustering coefficient
  - What portion of node i's neighbors are connected?
  - Node *i* with degree k_i
  - $C_i \in [0,1]$

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

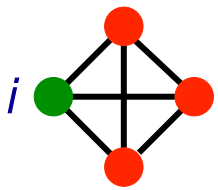
where e_i is the number of edges between the neighbors of node *i*



$C_i=0$



$C_i=1/3$



$C_i=1$

$$C_i = \frac{\text{\# of pairs of neighbors of } i \text{ that are connected}}{\text{\# of pairs of neighbors of } i}$$

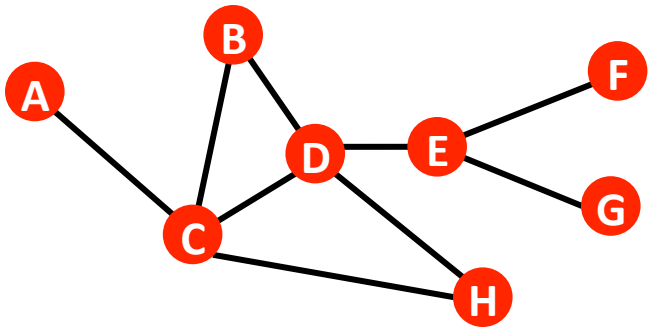
Average clustering coefficient: $C = \frac{1}{|V|}\sum_{i=1}^{|V|} C_i$

Clustering Coefficient (2/2)

- Clustering coefficient
  - What portion of node i's neighbors are connected?
  - Node *i* with degree k_i
  - $C_i \in [0,1]$

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

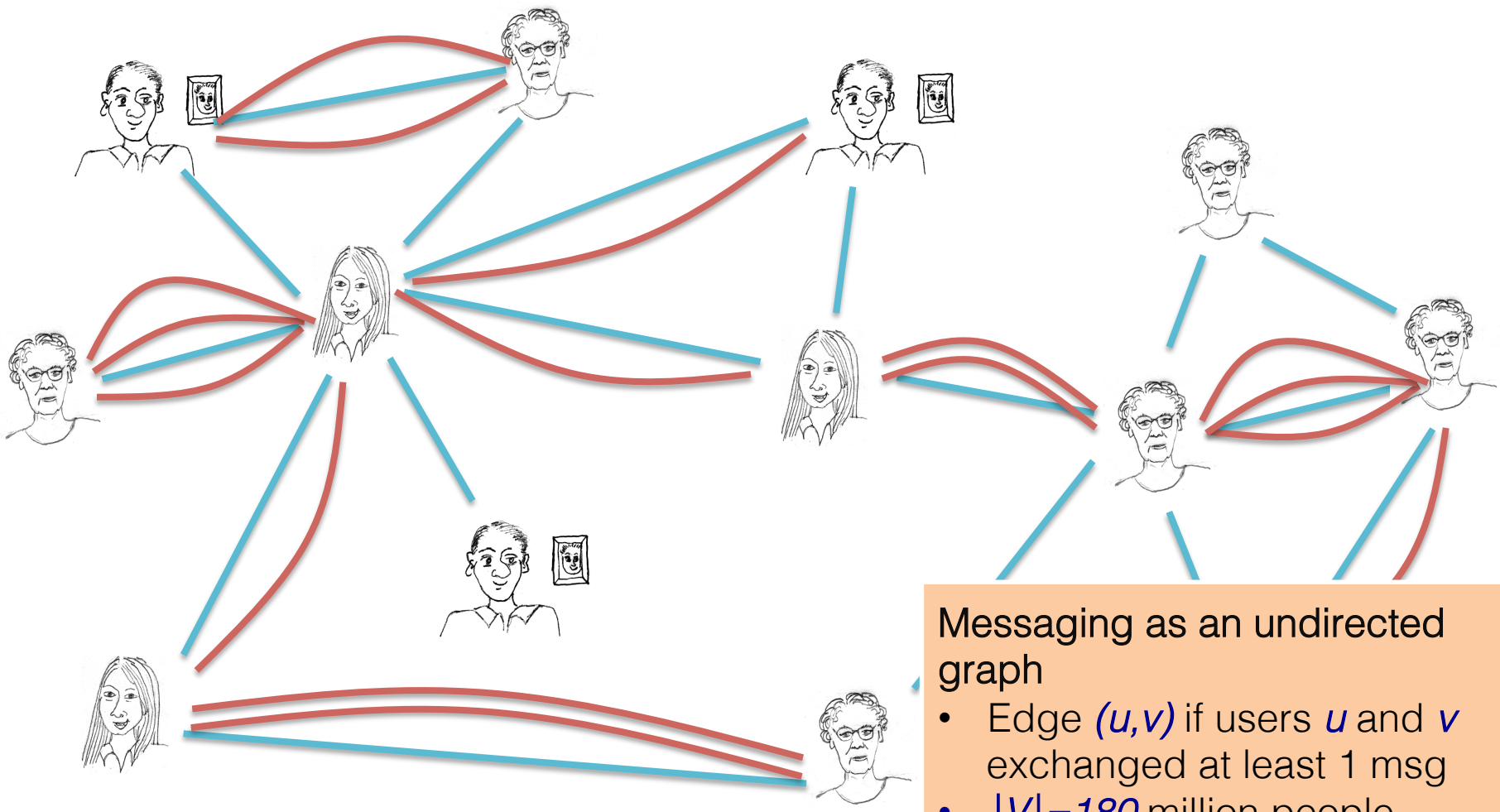
where e_i is the number of edges between the neighbors of node *i*



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

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

MSN: Messaging as a Multigraph



Messaging as an undirected graph

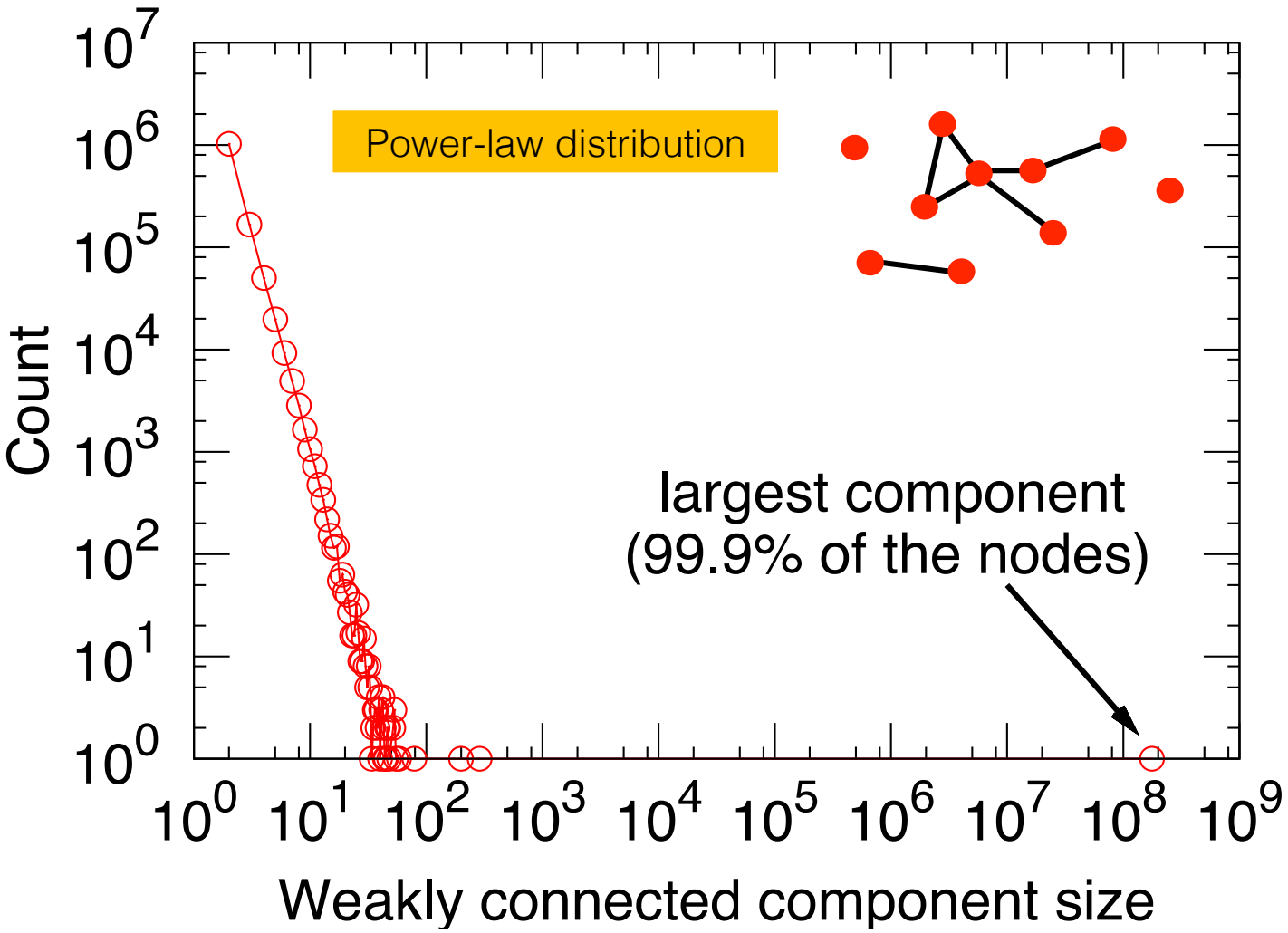
- Edge *(u,v)* if users *u* and *v* exchanged at least 1 msg
- $|V|=180$ million people
- $|E|=1.3$ billion edges

Three Basic Network Properties

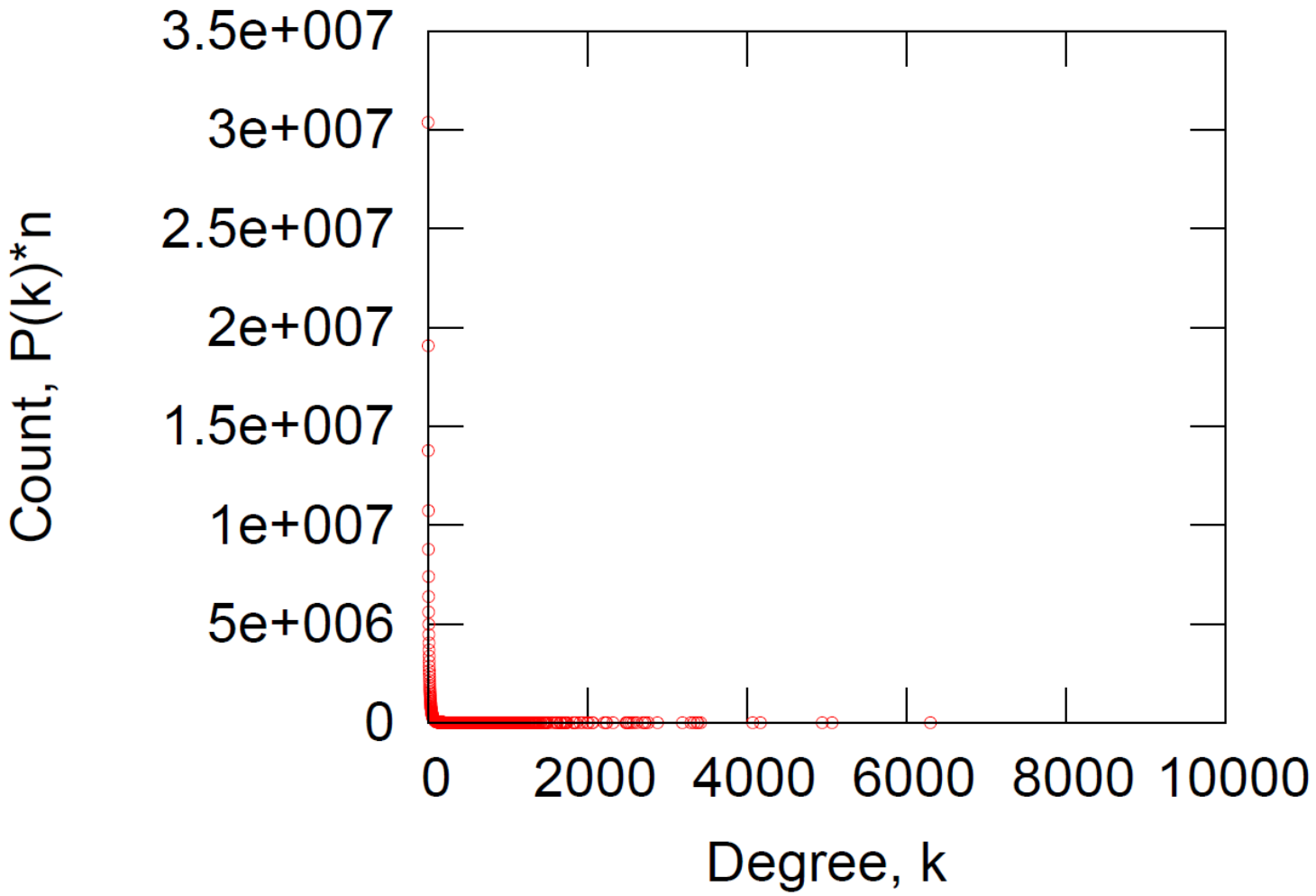
Degree distribution:	*P(k)*
Path length:	*h*
Clustering coefficient:	*C*

What did we observe in the MSN graph?

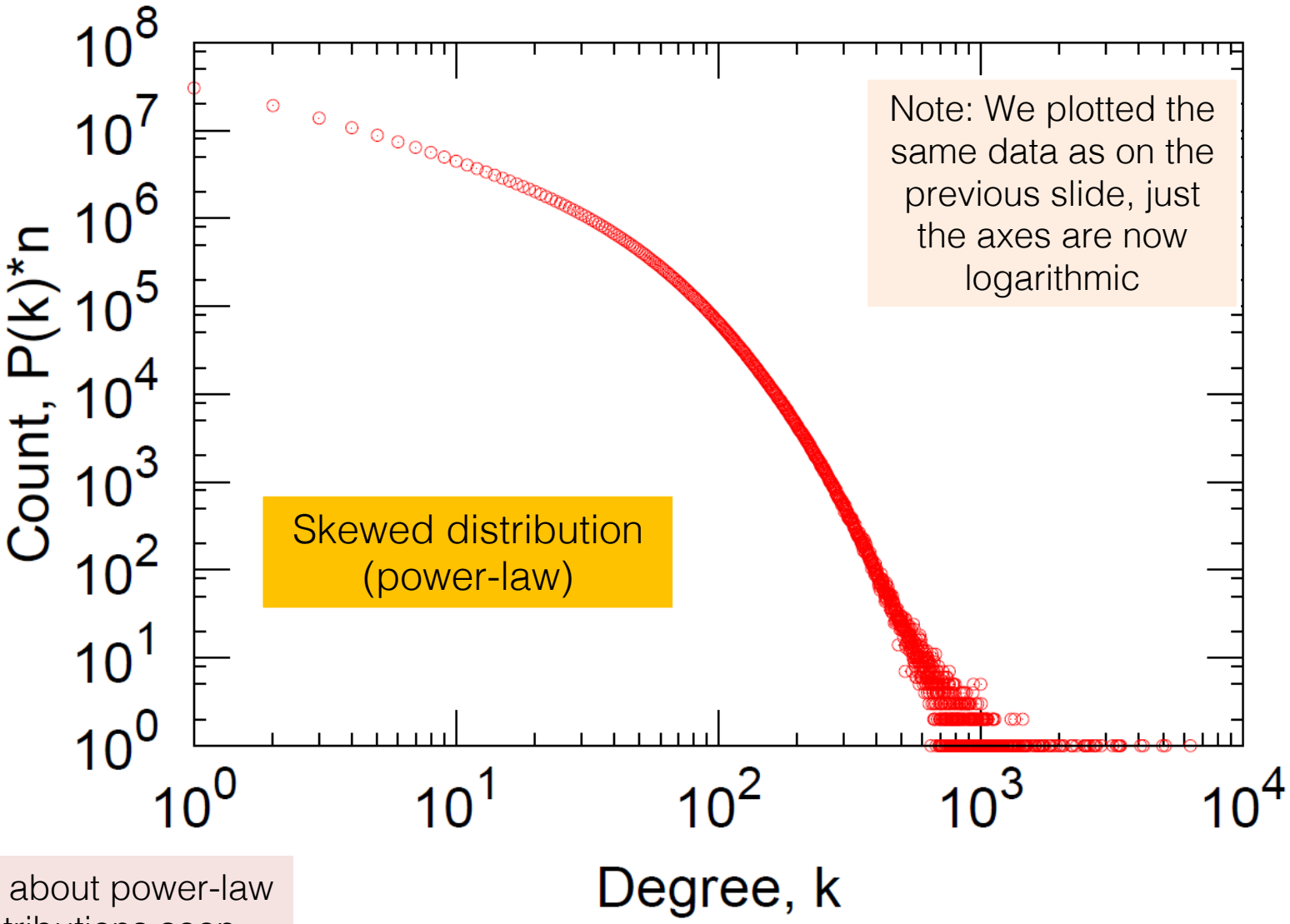
MSN: Connectivity



MSN: Degree Distribution



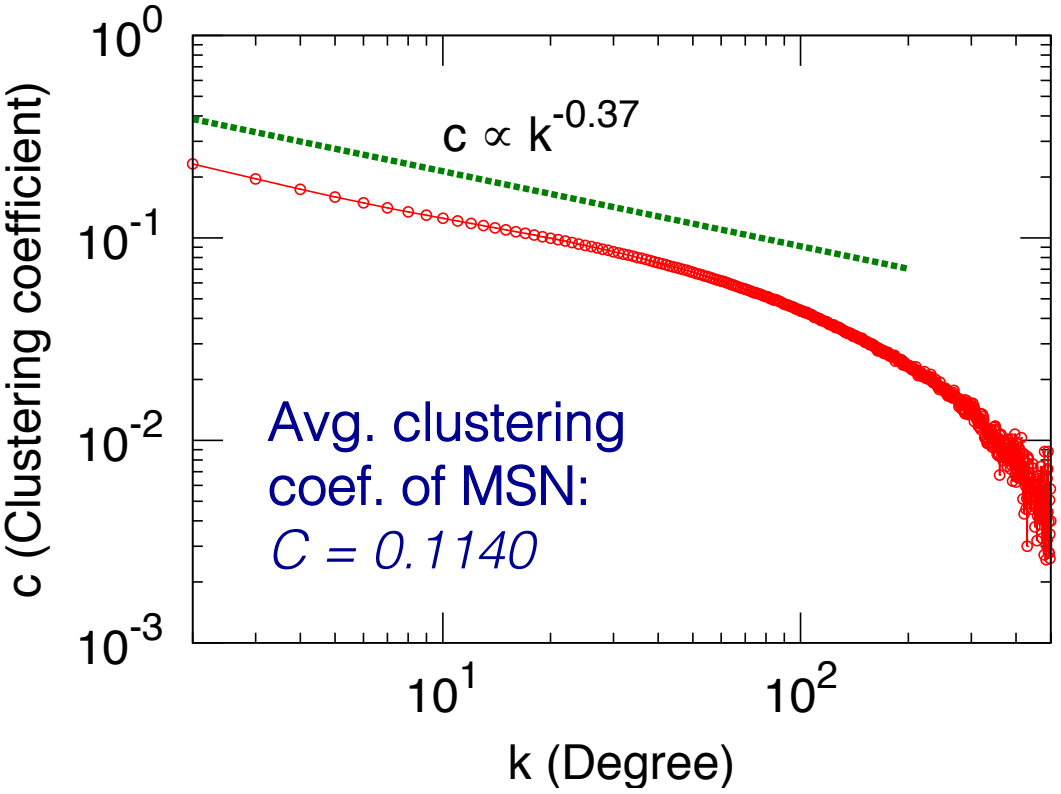
MSN: Log-Log Degree Distribution



More about power-law distributions soon

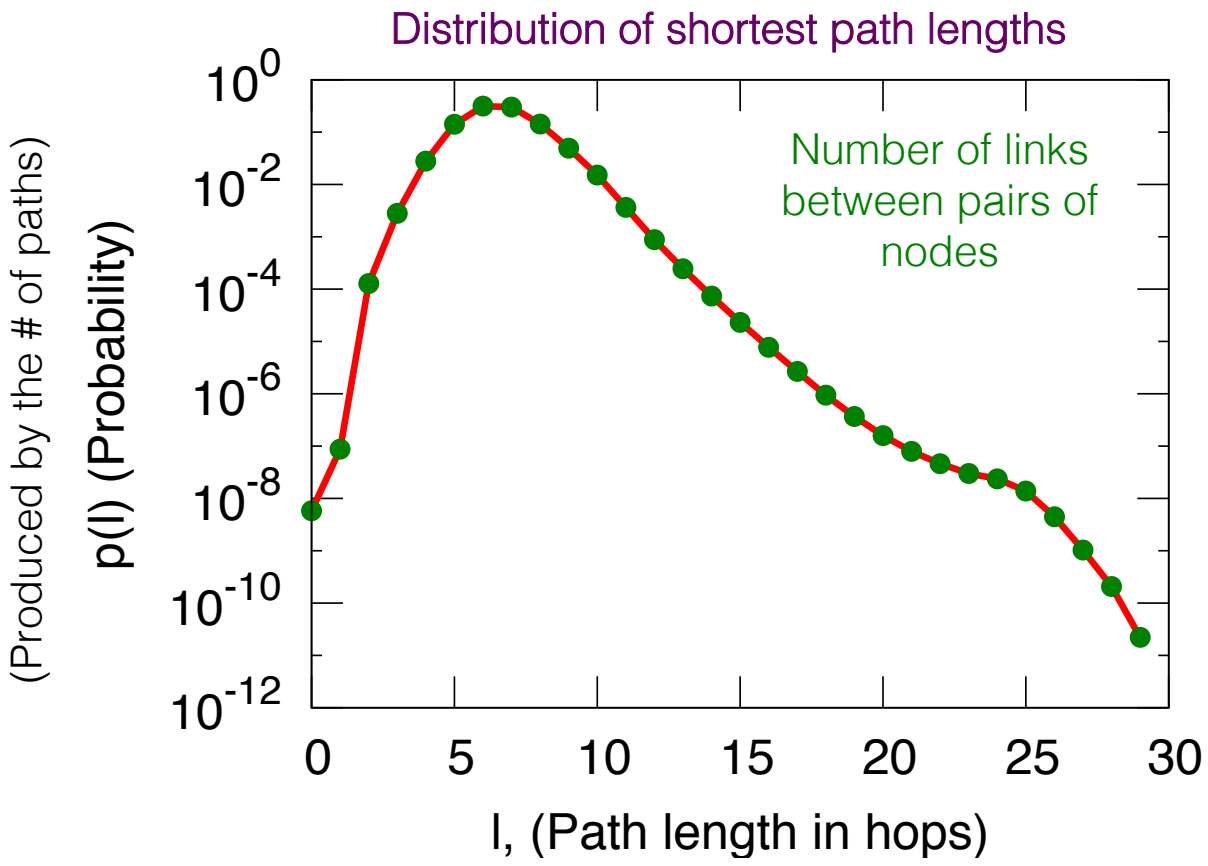
MSN: Clustering

Social networks are highly transitive: people with common friends tend to be friends themselves -> Use clustering coefficient to quantify this property



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

MSN: Diameter



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: Summary of Properties

Degree distribution:	*Heavily skewed* *avg. degree= 14.4*
Path length:	*6.6*
Clustering coefficient:	*0.11*

Are these values “expected”?
Are they “surprising”?

To answer this we need a model

Overview of this Lecture

- The Erdős–Rényi random graph model
  - Degree distribution and other properties
  - Does it capture the properties of real-networks?

- Small-world phenomenon

The Erdős–Rényi random graph model

Erdös-Rényi Random Graphs

[Erdös and Rényi, 60]

Two variants:

- $G_{n,p}$: undirected graph on n nodes and each edge (u, v) appears i.i.d. with probability p

our focus

- $G_{n,m}$: undirected graph with n nodes, and m uniformly at random picked edges

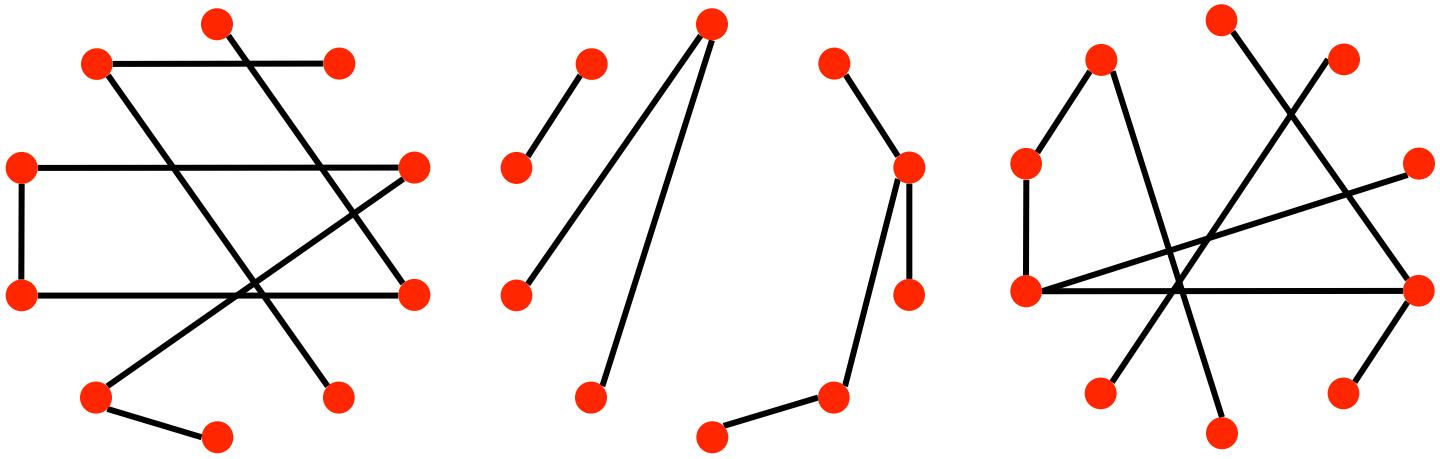
What kinds of networks does such model produce?

The Erdös-Rényi Random Graph Model

- In terms of the adjacency matrix

$$\forall_{i>j} \quad \mathbf{A}_{ij} = \mathbf{A}_{ji} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{otherwise} \end{cases}$$

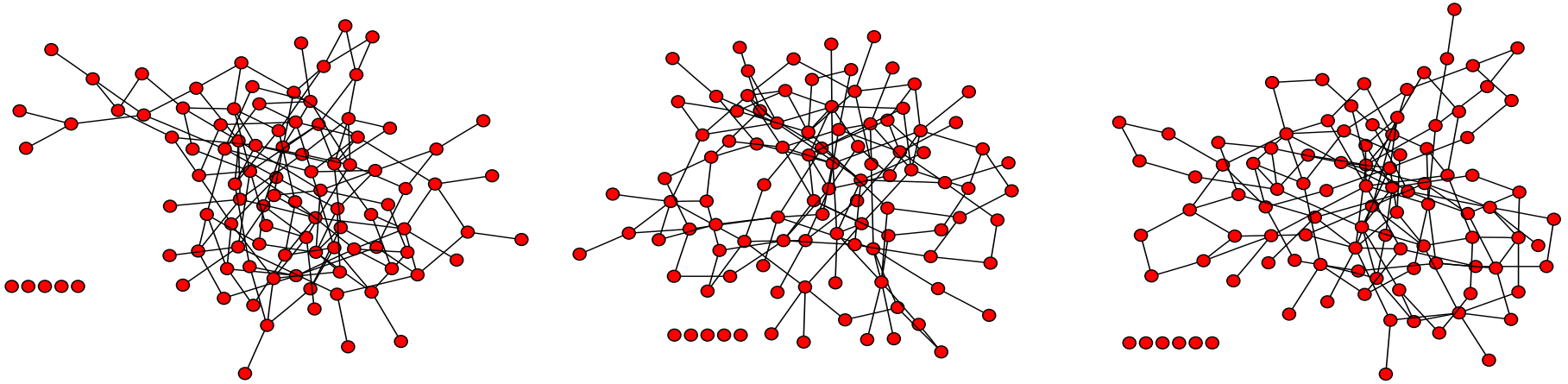
- Observation: *n* and *p* do not uniquely determine the graph
  - The graph is a result of a random process
- We can have many different realizations given the same *n* and *p*



n = 10
p = 1/6

Example of $G_{n,p}$

p= 0.03
n =100



Properties of $G_{n,p}$

Degree distribution:	$P(k)$
Path length:	h
Clustering coefficient:	C

What are the values of these properties for $G_{n,p}$?

Random Graph Model: Edges

- Question: How likely is a graph with *m* edges?

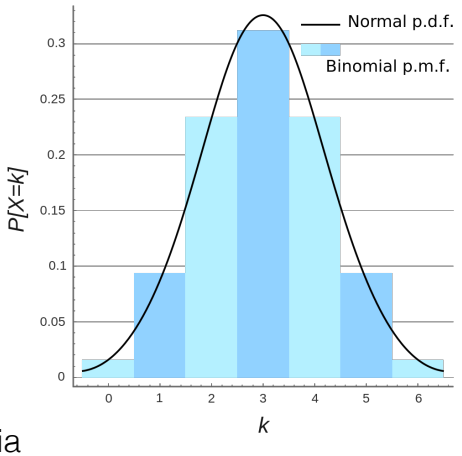
- The probability that a given $G_{n,p}$ generates a graph on exactly *m* edges:

$$\Pr(m) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m}$$

Notice that $\binom{n}{2}$ is the maximum number of edges (recall the complete graph)

Pr(m) is exactly the Binomial distribution

(Number of successes in a sequence of $\binom{n}{2}$ independent yes/no experiments)



Source: Wikipedia

What is the Average Number of Edges?

- The mean value <m> of edges can be derived as follows

$$\langle m \rangle = \sum_{m=0}^{\binom{n}{2}} m \ \Pr(m)$$

$$= \binom{n}{2} p$$

- Intuition: the **expected** number of edges in the graph is equal to the number of possible edges, given that each edge exists with probability *p*

Proof idea:
Use the *Binomial theorem*: If random variable $X \sim B(n,p)$, then $E[X] = np$
(linearity of expectations)

What is the Expected Degree?

- Recall that in a graph with *m* edges, the mean degree of a vertex is <k> = 2m/n
- In the $G_{n,p}$ model, the expected degree can be derived from *<m>* shown before

$$
\begin{aligned}
c = \langle k \rangle &= \sum_{m=0}^{\binom{n}{2}} \frac{2m}{n} \Pr(m) \\
&= \frac{2}{n} \binom{n}{2} p \\
&= (n-1)\, p
\end{aligned}
$$

Or: $\langle k \rangle = \frac{2\langle m \rangle}{n} = \frac{2}{n} \binom{n}{2} p = (n-1)\, p$

Degree Distribution (1/4)

- The degree distribution is binomial

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

Fraction of nodes with degree k

Select k nodes out of n-1

Probability of having k edges

Probability of missing the rest n-1-k edges

Mean and variance of a binomial distribution

$$\langle k \rangle = (n-1)p$$

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

What is happening when $n \to \infty$?

Degree Distribution (2/4)

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

When $n \to \infty$ then $p = c/(n-1)$ will be vanishingly small

Approximation

$$\ln\left[(1-p)^{n-1-k}\right] = (n-1-k)\ln\left(1-\frac{c}{n-1}\right)$$

$$\approx (n-1-k)\frac{-c}{n-1} \approx -c$$

$\ln(1+x) \approx x$ when x is small

Taking the exponential of both sides

$$(1-p)^{n-1-k} \approx e^{-c}$$

The degree distribution is now: $\Pr(k) = \binom{n-1}{k} p^k e^{-c}$

Degree Distribution (3/4)

$$\Pr(k) = \binom{n-1}{k} p^k e^{-c}$$

Further simplification

$$\binom{n-1}{k} = \frac{(n-1)!}{(n-1-k)!\,k!} \approx \frac{(n-1)^k}{k!}$$

Thus, when $n \to \infty$ the degree distribution becomes:

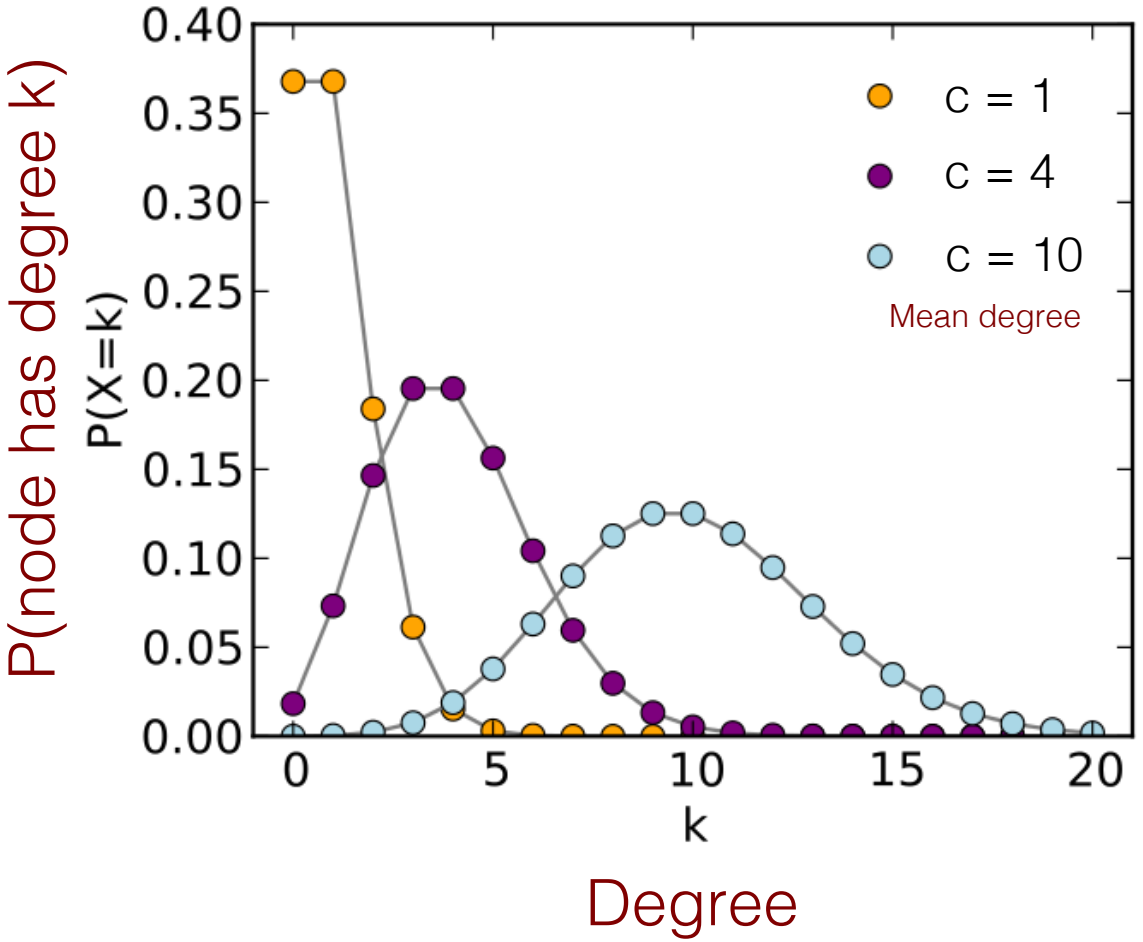
$$\Pr(k) \approx \frac{(n-1)^k}{k!} p^k e^{-c} = \frac{(n-1)^k}{k!} \left(\frac{c}{n-1}\right)^k e^{-c} = \frac{c^k}{k!} e^{-c}$$

Poisson distribution

Degree Distribution (4/4)

Poisson distribution

$$\Pr(k) = \frac{c^k}{k!} e^{-c}$$



Clustering Coefficient of $G_{n,p}$

- Recall that $C_i = \frac{2e_i}{k_i(k_i-1)}$

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

- Edges in $G_{n,p}$ appear with probability p

$$e_i = p\binom{k_i}{2} = p\frac{k_i(k_i-1)}{2}$$

Each pair is connected with prob. p

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

Thus,

$$C_i = \frac{pk_i(k_i-1)}{k_i(k_i-1)} = p$$

Clustering coefficient of $G_{n,p}$

$$C = \frac{1}{n}\sum_{i\in V} C_i = p = \frac{c}{n-1} \approx \frac{c}{n}$$

Clustering coefficient of a random graph is small
For a fixed avg. degree (that is $p=1/n$), C decreases with the graph size n

Properties of $G_{n,p}$

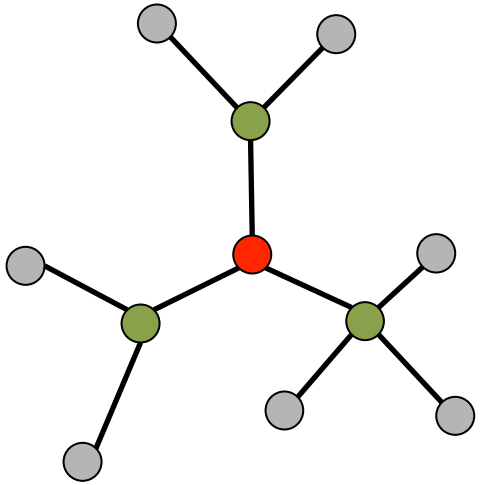
Degree distribution: $\Pr(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$

Clustering coefficient: $C = p \approx \frac{\langle k \rangle}{n}$

Path length: Next!

Diameter and Avg. Path Length

Random graphs tend to have a tree-like topology with almost constant node degrees



Random graphs have tree-like topology

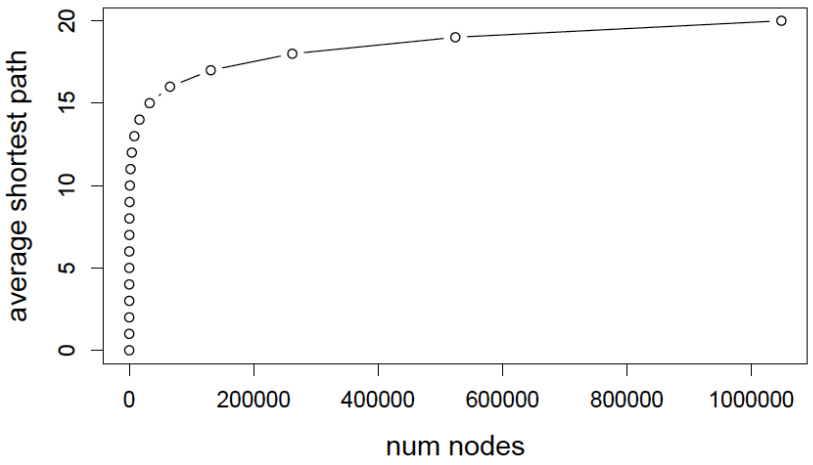
c: avg. degree

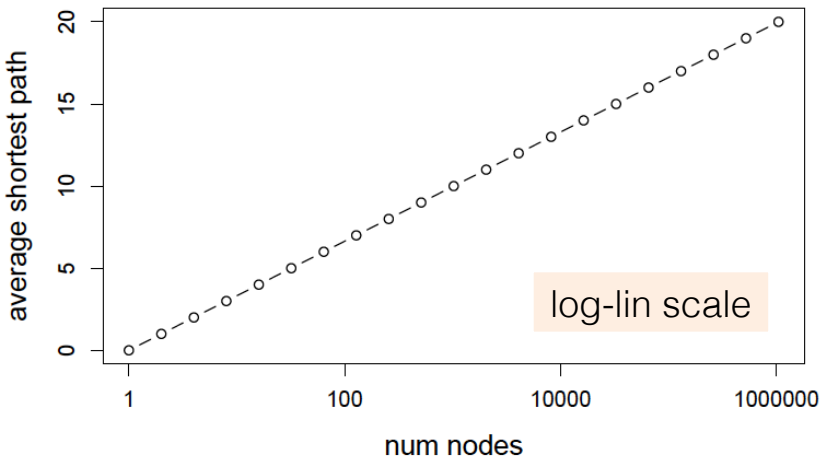
c nodes at distance one (*h=1*)
c^2 nodes at distance two (*h=2*)
c^3 nodes at distance three (*h=3*)
…
c^h nodes at distance *h*

$$n = 1 + c + c^2 + \ldots + c^{h_{\max}} = \frac{c^{h_{\max}+1} - 1}{c - 1} \approx c^{h_{\max}} \Rightarrow h_{\max} = \frac{\log n}{\log c} = O(\log n)$$

Diameter

Erdös-Rényi Average Shortest Path





Erdös-Rényi random graphs can grow to be very large, but nodes will be a few hops apart

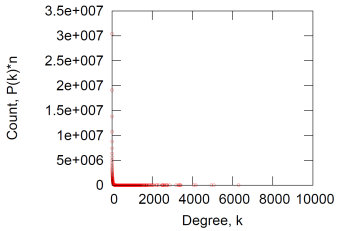
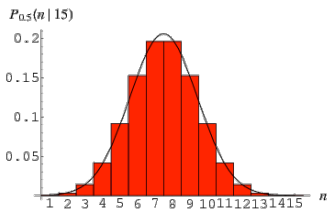
Properties of $G_{n,p}$

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

Path length: $O(\log n)$

Clustering coefficient: $C = p \approx \frac{\langle k \rangle}{n}$

MSN vs. $G_{n,p}$

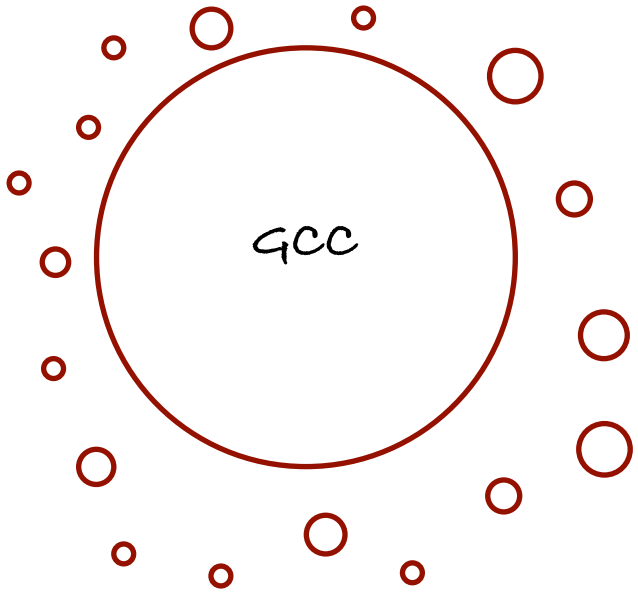
	MSN	$G_{n,p}$
Degree distribution:		 n=180*M*
Path length:	6.6	8.2 $O(\log n)$
Clustering coefficient:	0.11	$\approx 8 \cdot 10^{-8}$ $C = p \approx \frac{\langle k \rangle}{n}$

Connected Components

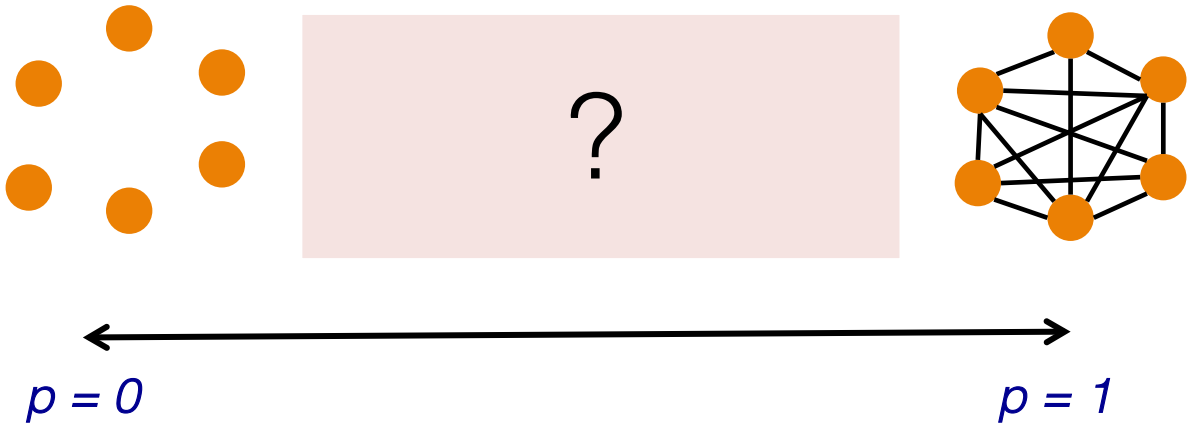
- Recall that, in real networks
  - There is a large number of connected components
  - The majority of the nodes belong the the largest (giant) connected component

Network	$\|V\|$	$\|E\|$	S
WWW (Alta Vista)	203 *M*	1.5 *B*	0.914
Facebook (Ugander et al. 2011)	721 *M*	68 *B*	0.99
Email messages (Newman)	59 *K*	86 *K*	0.95
Co-authorships (Biology)	1.5 *M*	12 *M*	0.92

S : fraction of nodes in the largest component



Connectivity of $G_{n,p}$



- Connectivity of random graphs depends on *p*
- Property: sudden appearance of a giant component as we vary the average degree *c* (recall that *c = (n-1)p*)
  - This phenomenon is called phase transition

Back to Node Degrees of $G_{n,p}$

- To have constant degree, let *p=c/(n-1)*
- Observation: If we build random graph $G_{n,p}$ with very small *p=c/(n-1)* we have many isolated nodes
- Why is this happening?

$$P[v \text{ has degree } 0] = (1-p)^{n-1} = \left(1 - \frac{c}{n-1}\right)^{n-1} \xrightarrow[n\to\infty]{} e^{-c}$$

How?

$$\lim_{n\to\infty}\left(1-\frac{c}{n-1}\right)^{n-1} = \left(1-\frac{1}{x}\right)^{x\cdot c} = \left[\underbrace{\lim_{x\to\infty}\left(1-\frac{1}{x}\right)^{-x}}_{e}\right]^{-c} = e^{-c}$$

Use substitution $\frac{1}{x} = \frac{k}{n-1}$

$$e = \lim_{x\to\infty}\left(1+\frac{1}{x}\right)^{x}$$

No Isolated Nodes

- How big do we have to make *p* before we are likely to have no isolated nodes?
- We know that *P[v has degree 0] = e^{-c}*

- Event we are asking about is:
  - *I* = some node is isolated
  - $I = \bigcup_{v \in V} I_v$ where I_v is the event that *v* is isolated

- We have

$$P(I) = P\left(\bigcup_{v \in V} I_v\right) \leq \sum_{v \in V} P(I_v) = ne^{-c}$$

union bound

A_i

$$\left|\bigcup_i A_i\right| \leq \sum_i |A_i|$$

No Isolated Nodes

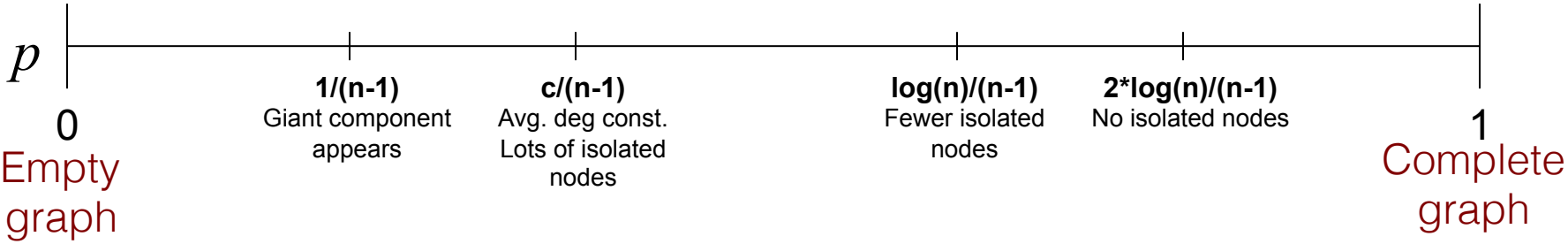
- We just learned: $P(I) \leq n\, e^{-c}$

- Let's try:
  - $c = \ln n$ then: $n\, e^{-c} = n\, e^{-\ln n} = n \cdot 1/n = 1$
  - $c = 2 \ln n$ then: $n\, e^{-2 \ln n} = n \cdot 1/n^2 = 1/n$

- Thus, if:
  - $c = \ln n$ then: $P(I) \leq 1$
  - $c = 2 \ln n$ then: $P(I) \leq 1/n \rightarrow 0$ as $n \rightarrow \infty$

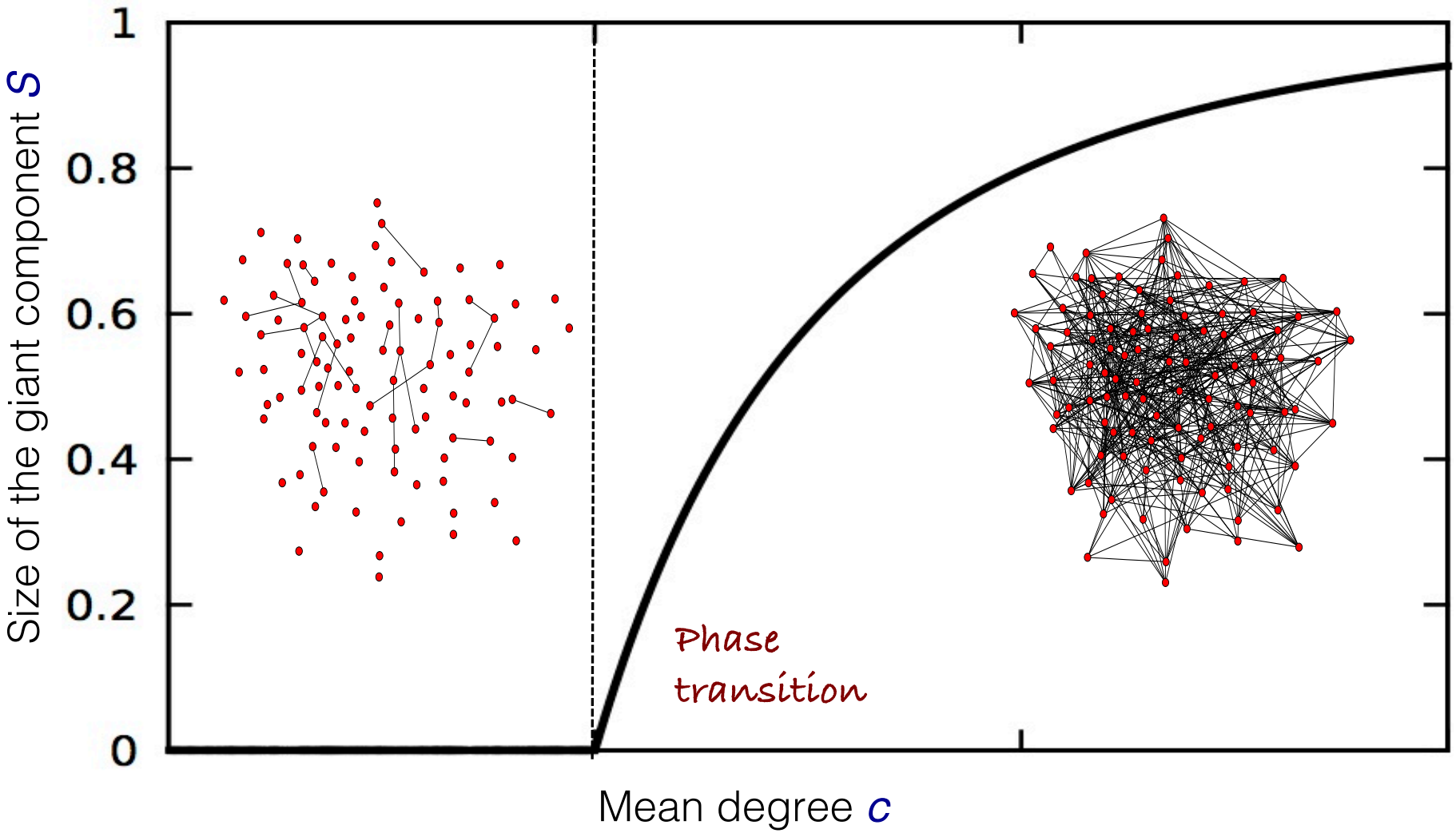
So, for $p = 2\ln(n)$ we get no isolated nodes (as $n \rightarrow \infty$)

“Evolution” of a Random Graph (1/3)

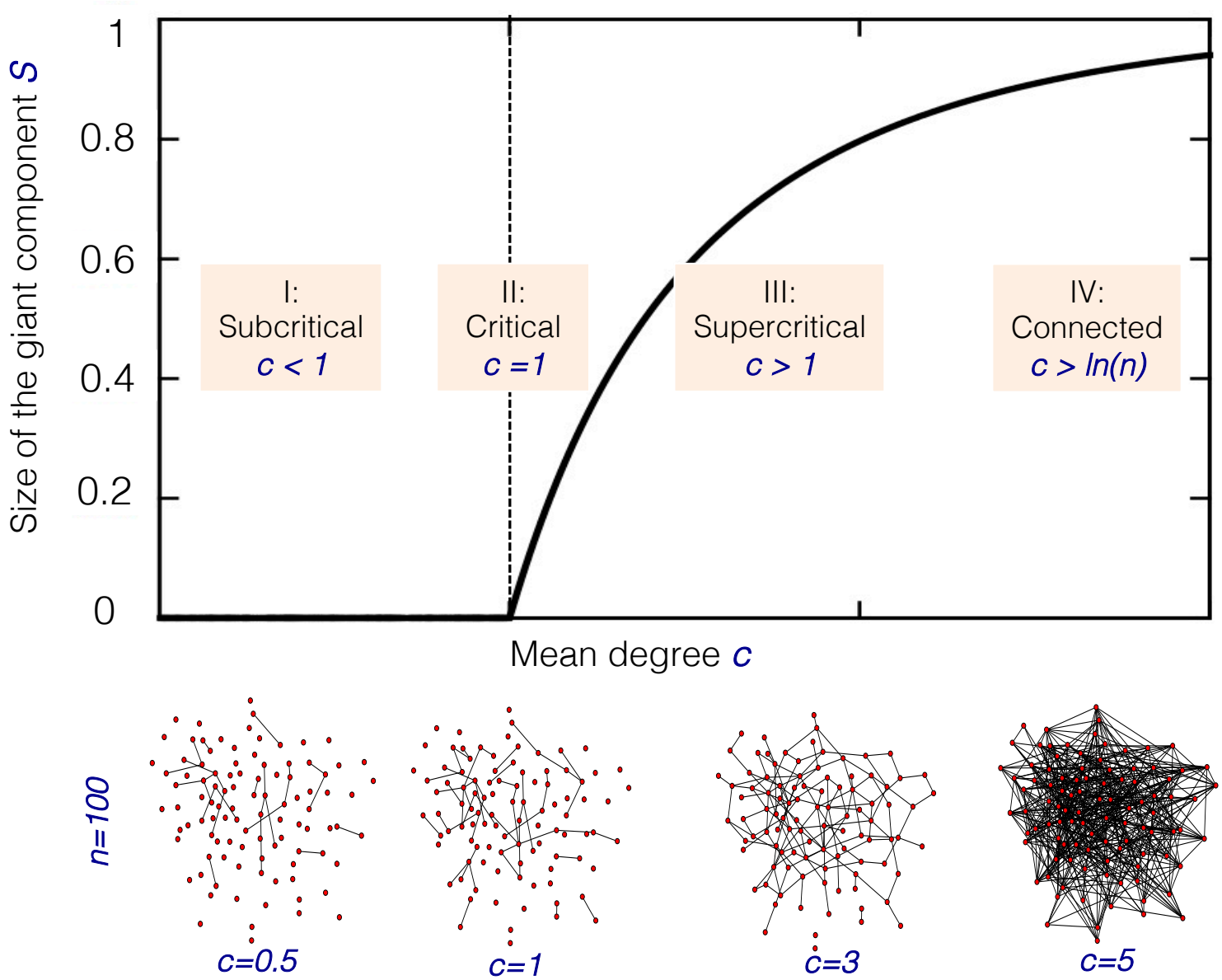


Graph structure of $G_{n,p}$ as p changes

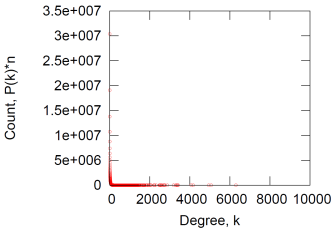
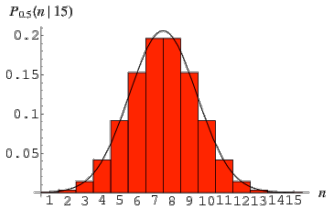
“Evolution” of a Random Graph (2/3)



“Evolution” of a Random Graph (3/3)



MSN vs. $G_{n,p}$

	MSN	$G_{n,p}$ n=180*M*	
Degree distribution:			✗
Path length:	6.6	8.2 $O(\log n)$	✓
Clustering coefficient:	0.11	$\approx 8 \cdot 10^{-8}$ $C = p \approx \frac{\langle k \rangle}{n}$	✗
Conn. Component:	0.99	GCC exists when *c>1*	✓

Real Networks vs. $G_{n,p}$

- Are real networks like random graphs?
  - Giant connected component: ☺
  - Average path length: ☺
  - Clustering Coefficient: ☹
  - Degree Distribution: ☹
- Problems with the random graph model
  - The degree distribution differs from that of real networks
  - The 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!*

The Epilogue!

- If $G_{n,p}$ is wrong, why did we spend time on it?
  - It is a reference model in network science
  - It will help us calculate many quantities, that can then be compared to the real data
  - It will help us understand to what degree a particular property is the result of some random process
  - Note: random graph theory was never meant to serve as a model of real systems

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

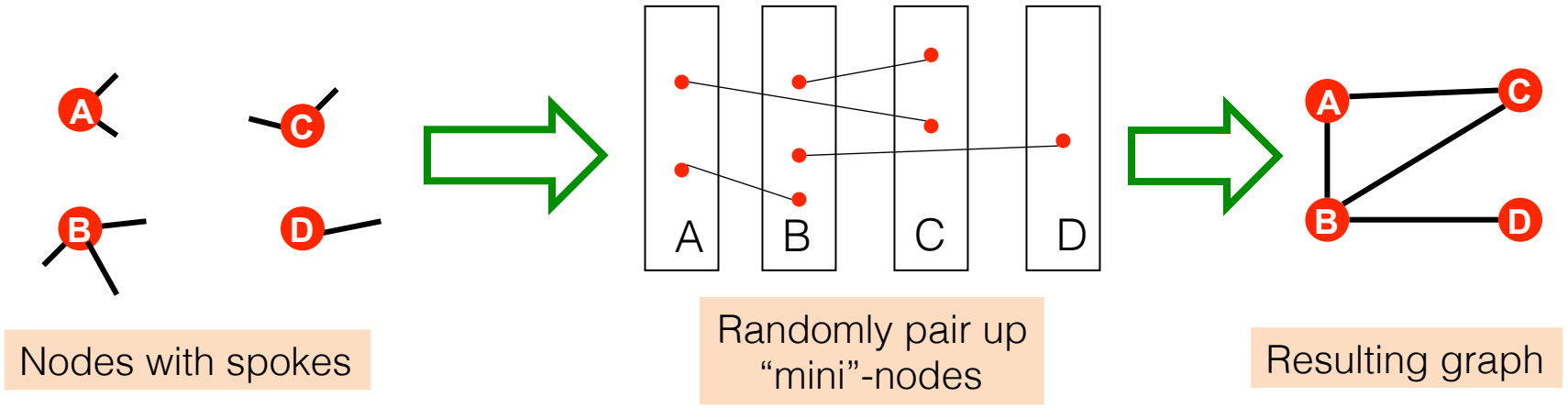
Evolution of the $G_{n,p}$ Graph

Time for a short video:

https://www.youtube.com/watch?v=mpe44sTSoF8

Configuration Model

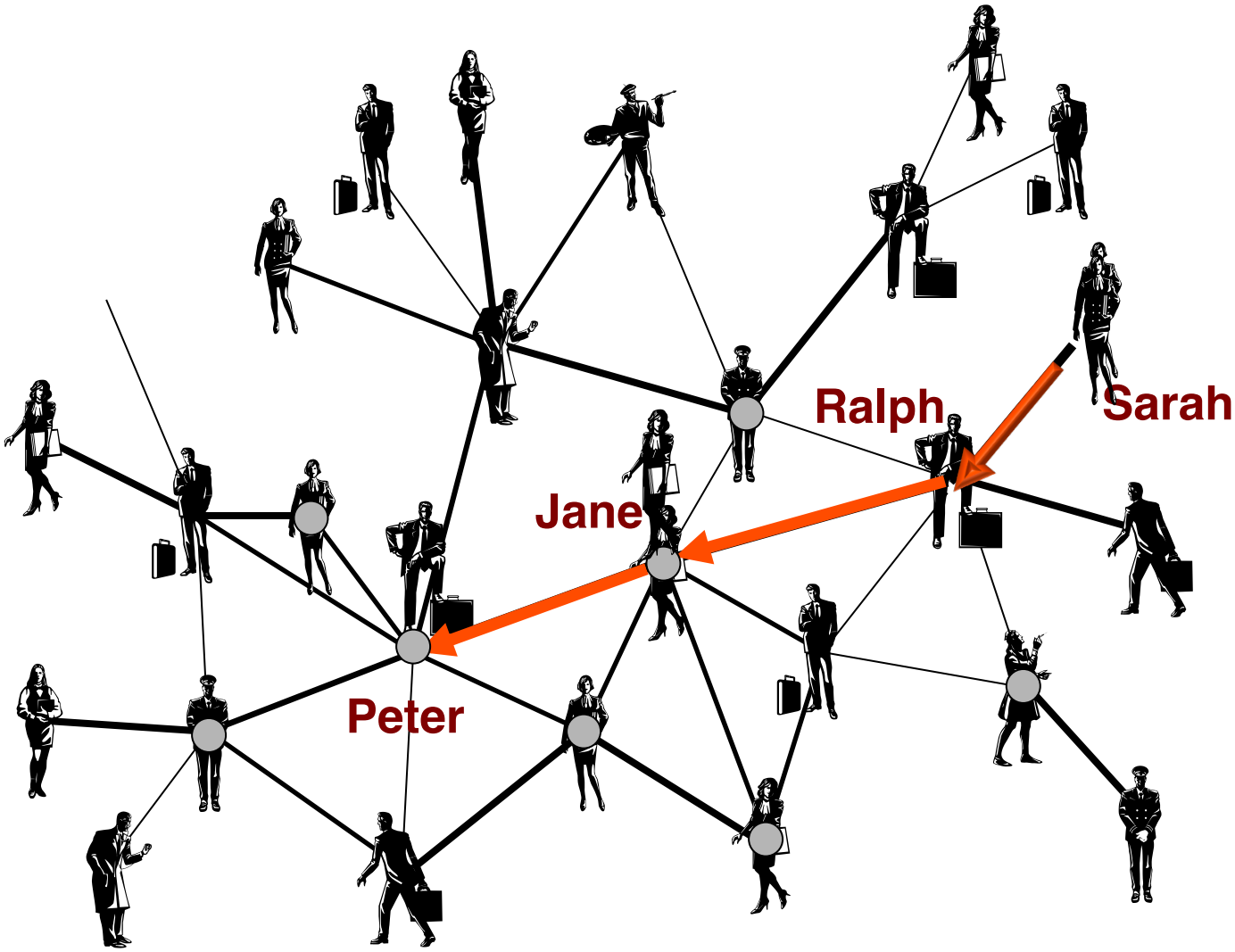
- Goal: Generate a random graph with a given degree sequence $k_1, k_2, \ldots k_n$
- Idea: use a configuration model:



- 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*

Small-world model

It's a Small World



Frigyes Karinthy, 1929
Stanley Milgram, 1967

The Small-World Phenomenon (1/2)

- Question: What is the typical shortest path length between any two people?

Stanley Milgram

- Small-world experiment (1967)
  - Randomly selected people in Nebraska were asked to send letters to Boston, by contacting somebody with whom they had direct connection
  1. People either sent the letter directly to the recipient
  2. Or to somebody they believed had a high likelihood of knowing the target

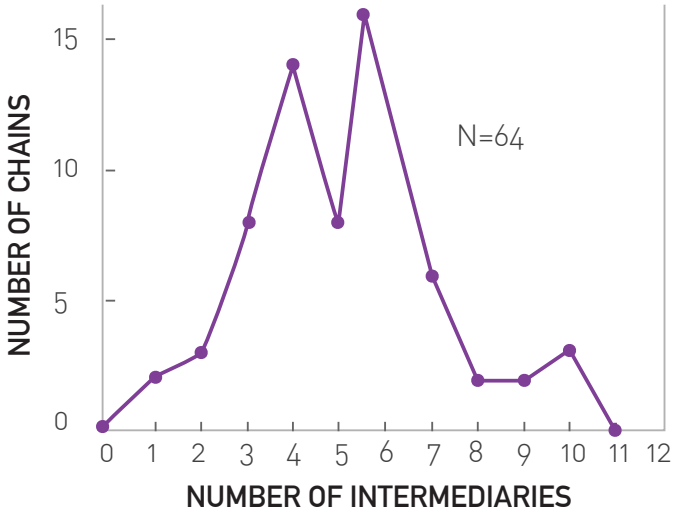
How many steps did it take?

The Small-World Phenomenon (2/2)

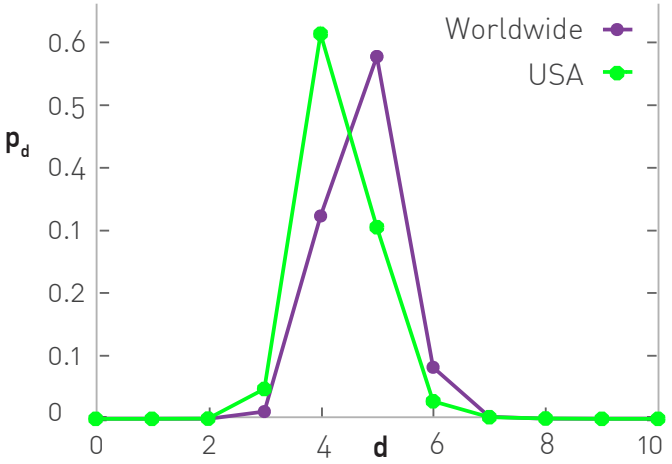
- For those letters that reached their destination, the average path length was 5.5 to 6
  - 64 messages reached their destination
  - Starting points and the target were non-random

- Facebook experiment (2011)
  - 721*M* nodes and 68*B* edges
  - Average distance between the users was 4.74

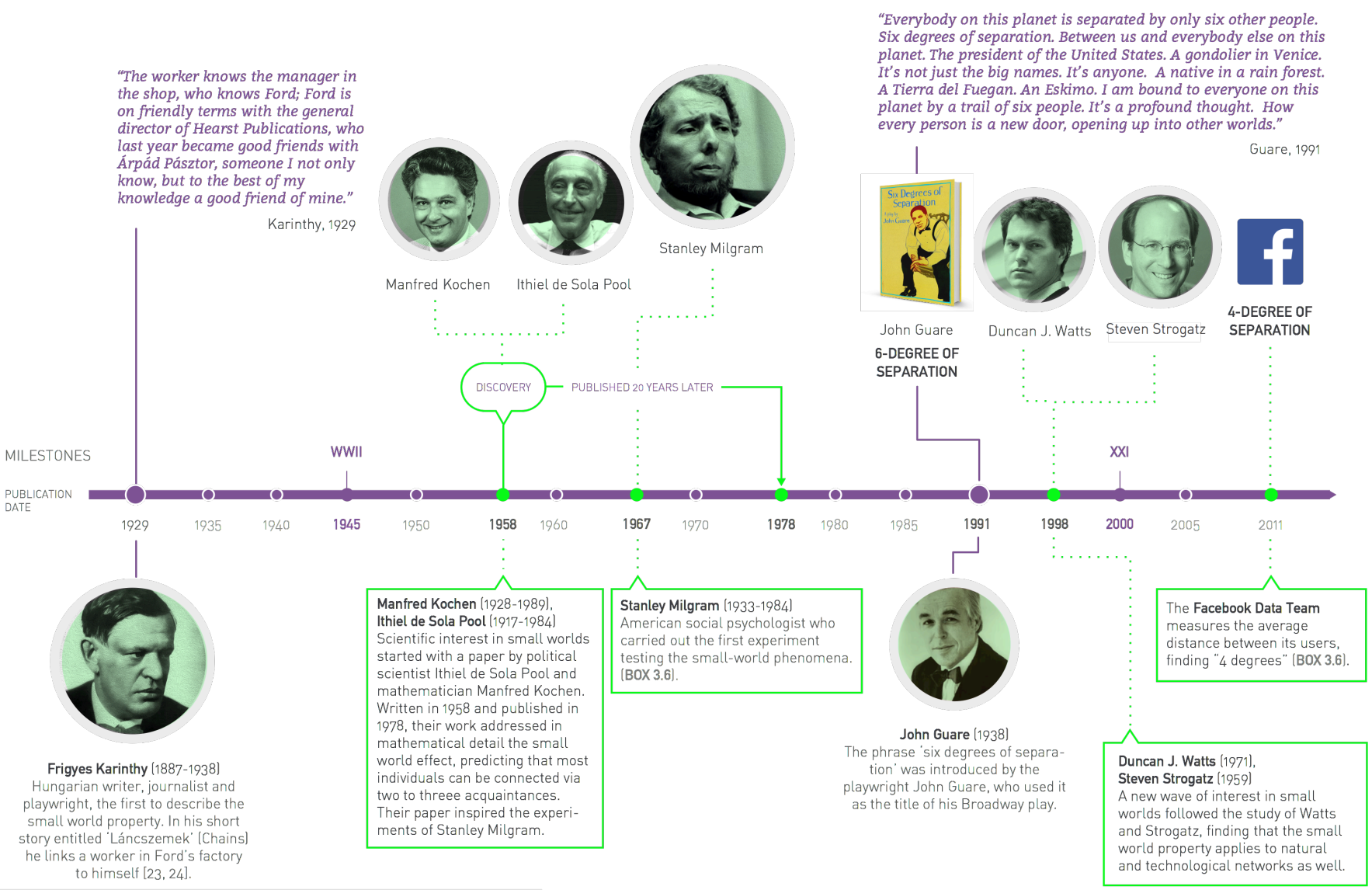
Milgram's small world experiment



Facebook small world experiment



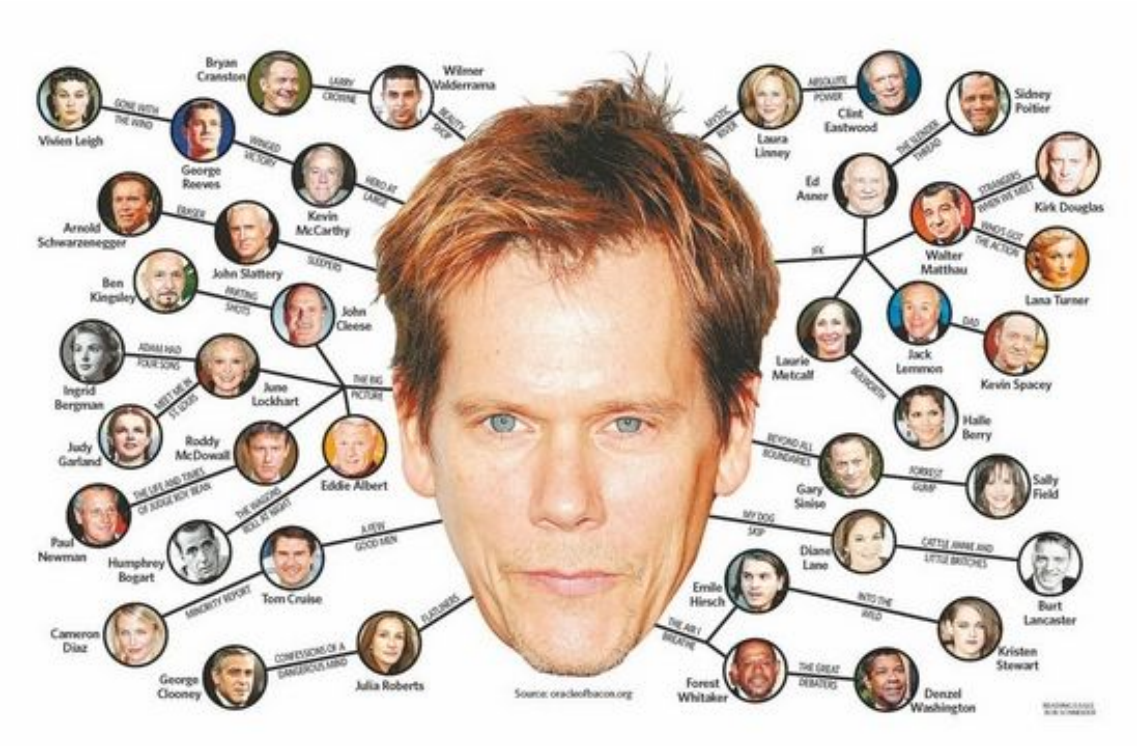
Three, Four or Six Degrees of Separation?



Source: http://barabasi.com/networksciencebook

Six Degrees of Kevin Bacon

The small-world phenomenon appears in various network settings

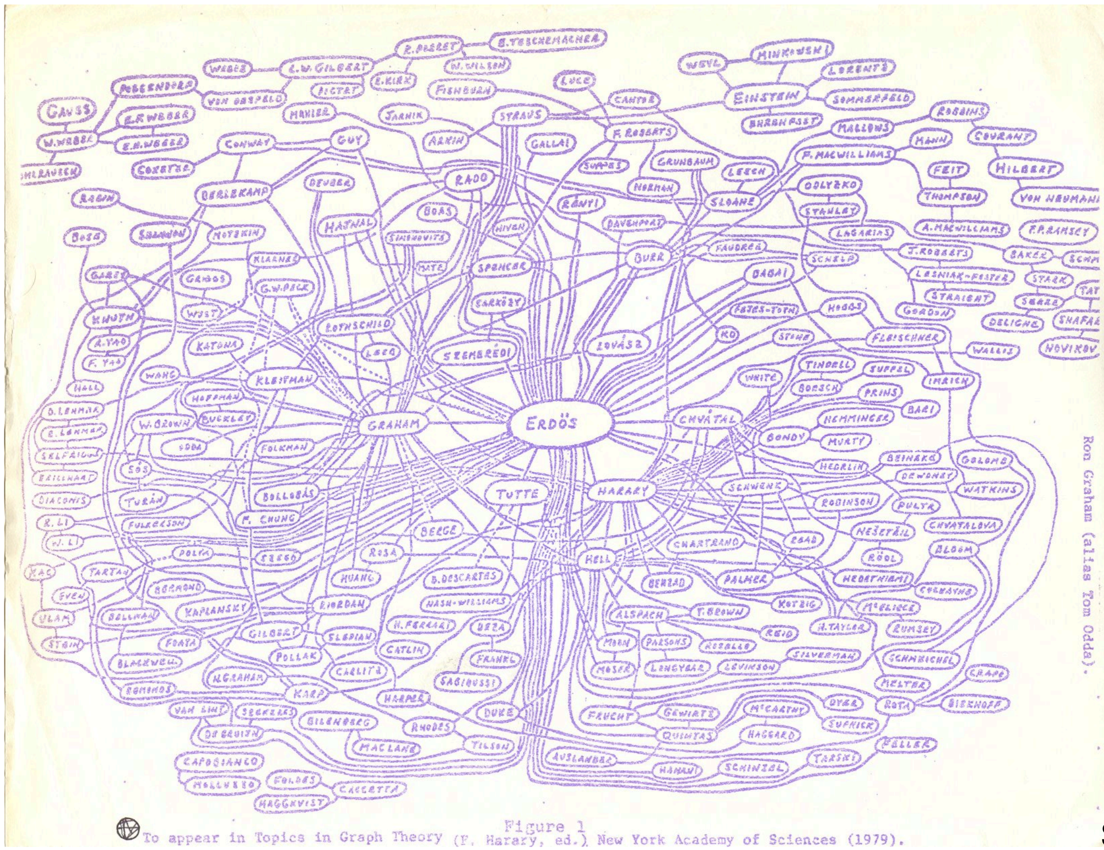


Source: www.classtools.net

Bacon number of an actor: number of degrees of separation from K. Bacon

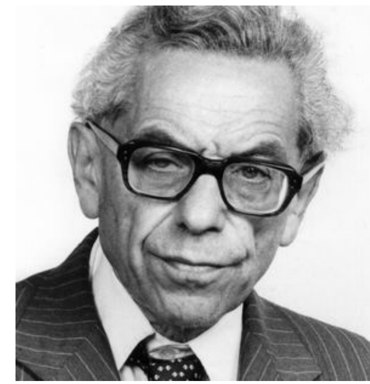
Erdős Numbers

The small-world phenomenon appears in various network settings



Source: physicsbuzz.physicscentral.com

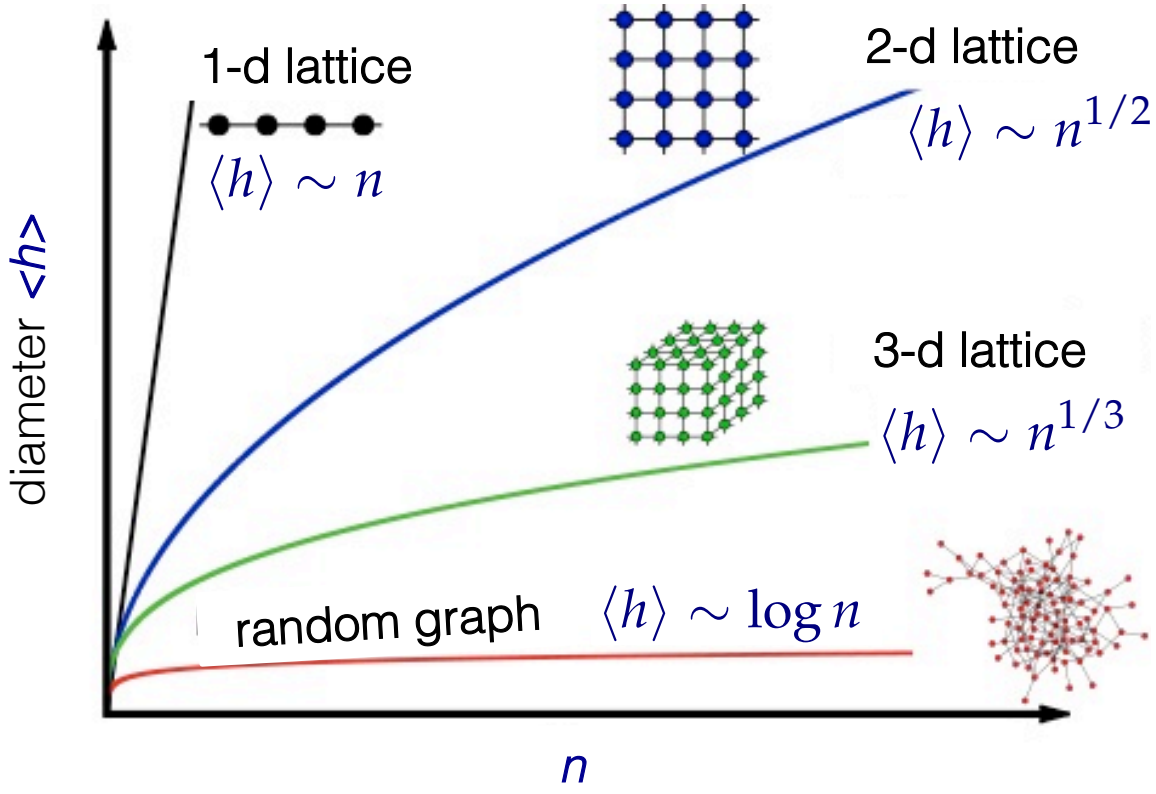
Paul Erdős



Source: UCSD

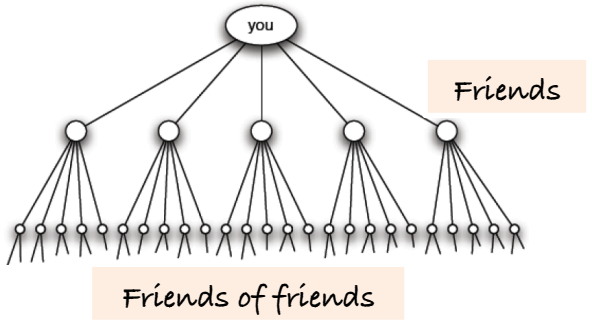
Erdős number: # of hops needed to connect the author of a paper to Paul Erdős

Our Intuition about Distance

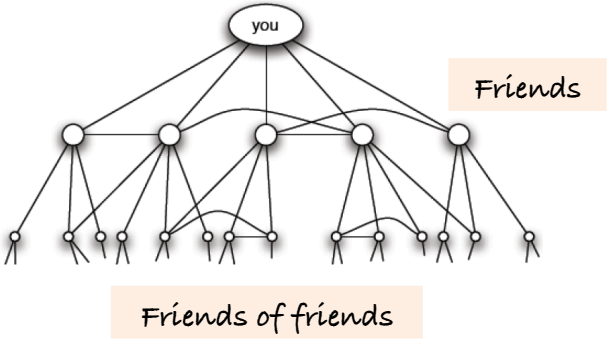


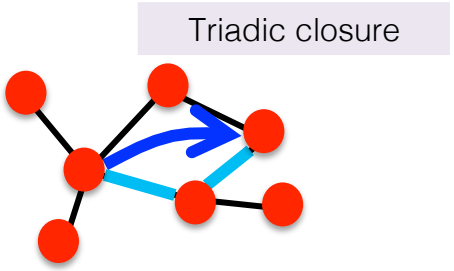
6-Degrees: Should We Be Surprised?

- 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?
  - 92% of new FB friendships are to a friend-of-a-friend [Backstom and Leskovec '11]





Clustering Implies Edge Locality

- MSN network has 7 orders of magnitude larger clustering than the corresponding $G_{n,p}$!

- Other examples:

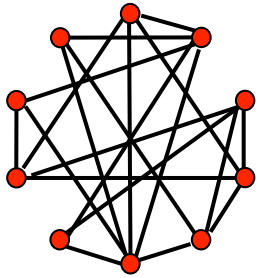
Actor Collaborations (IMDB): *n = 225,226* nodes, avg. degree *<k> = 61*
Electrical power grid: *n = 4,941* nodes, *<k>= 2.67*
Network of neurons: *n = 282* nodes, *<k> = 14*

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

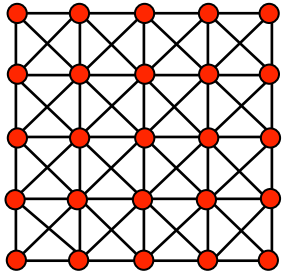
The "Controversy"

- Random graphs
  - Short avg. path lenth: *O(log n)*
    - This is "best" we can do if we have a constant degree
  - But clustering is low!



Low diameter
Low clustering coefficient

- But networks have "local" structure
  - Triadic closure:
    Friend of a friend is my friend
  - High clustering but
    diameter is also high

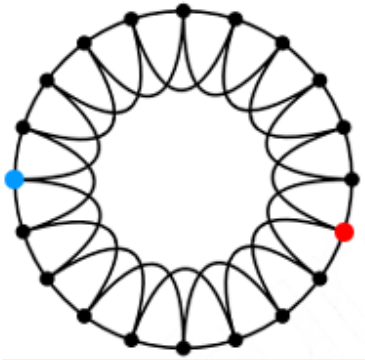


High clustering coefficient
High diameter

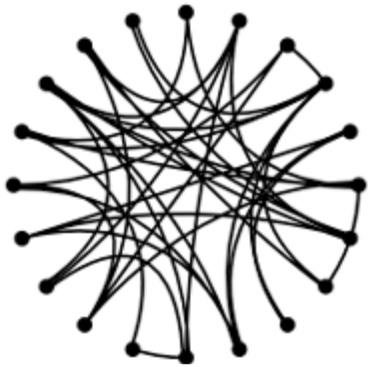
How can we have both?

Small-World: How?

- Could a network with high clustering be at the same time a small world?
  - How can we have at the same time **high clustering** and **small diameter?**



High clustering
High diameter



Low clustering
Low diameter

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

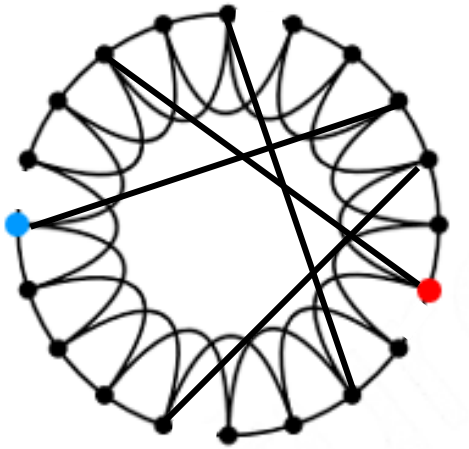
Solution: The Small-World Model

Small-world Model [Watts, Strogatz '98]

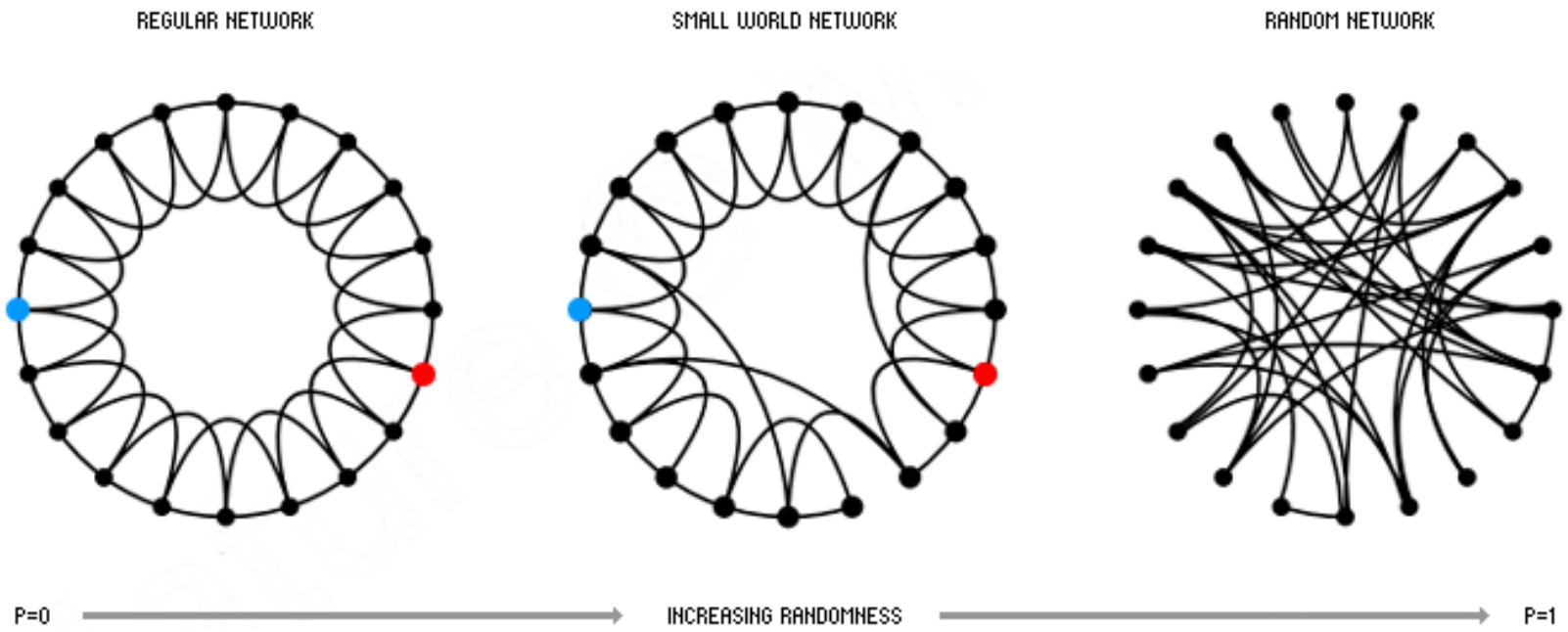
Two components of the model:

- (1) Start with a low-dimensional regular lattice
  - (In our case we 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

Reduce diameter



The Small-World Model



High clustering
High diameter

$$h = \frac{n}{2\bar{k}} \qquad C = \frac{3}{4}$$

High clustering
Low diameter

Low clustering
Low diameter

$$h = \frac{\log n}{\log \bar{k}} \qquad C = \frac{\bar{k}}{n}$$

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

Real network are not random

They follow properties that
deviate from randomness

Next Part of Today's Lecture

- Power-law degree distribution of real networks
- Preferential Attachment model
- Consequences of skewed degree distribution
  - Robustness of network

Thank You!