The Random Network Model

Alberto Paccanaro
EMAp – FGV

www.paccanarolab.org

Some material and images are from (or adapted from):
A. Barabási, and M. Pósfai. Network science, Cambridge University Press, 2016

Random Network Theory

- We want to build models that reproduce the properties of real networks

- Most networks look as if they are random (but they are not)

- Random network theory constructs and characterizes networks that are truly random.

Random Networks

A random network, *G(N, p)* consists of *N* nodes where each node pair is connected with probability *p*.

To construct a random network:

1) Start with N isolated nodes.
2) Select a node pair and connected it with probability p

   (generate a random number between 0 and 1. If the number exceeds p, connect the node pair with a link, otherwise leave the nodes disconnected).

3) Repeat step (2) for each of the N(N-1)/2 node pairs.

- **Also known as Erdős-Rényi networks**

Binomial distribution

Bernoulli distribution: discrete probability distribution of a random variable which takes the value *1* with probability *p* and the value *0* with probability *(1-p)*.

Prob mass function $p^k\,(1-p)^{1-k}$ for $k \in \{0,1\}$

Binomial distribution: discrete probability distribution of the number of *1 (successes)* in a sequence of n independent experiments where 1 has with probability p and 0 probability 1 – p (*prob of getting k successes in n trials*)

Prob mass function $\binom{n}{k}\, p^k (1-p)^{n-k}$

Number of links

The probability that a random network has exactly L links has a **binomial distribution**:

$$p_L = \binom{\frac{N(N-1)}{2}}{L} p^L (1-p)^{\frac{N(N-1)}{2}-L}$$

number of different ways we can place L links among N(N-1)/2 node pairs

probability that L of the attempts to connect the N(N-1)/2 pairs of nodes have resulted in a link

probability that the remaining N(N-1)/2 - L attempts have not resulted in a link

The binomial distribution describes the number of successes in N independent experiments with two possible outcomes, in which the probability of one outcome is p, (and of the other is 1-p)

Expected number of links:

$$\boxed{\langle L \rangle = p \ \frac{N(N-1)}{2}}$$

$\langle L \rangle$ is the product of the probability p that two nodes are connected and the number of pairs we attempt to connect

Average degree

$$\boxed{\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1)}$$

$\langle k \rangle$ is the product of the probability p that two nodes are connected and *(N-1)*, which is the maximum number of links a node can have in a network of size N.

In a Random Network, if we increase *p:*

- $\langle L \rangle$ increases *linearly* from 0 to N(N-1)/2
- $\langle k \rangle$ increases *linearly* from 0 to N-1

Degree distribution

p_k : the probability that a randomly chosen node has degree k

The probability that node i has exactly k links has a **binomial distribution**:

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

number of ways we can select k links from N- 1 potential links a node can have

probability that k of its links are present

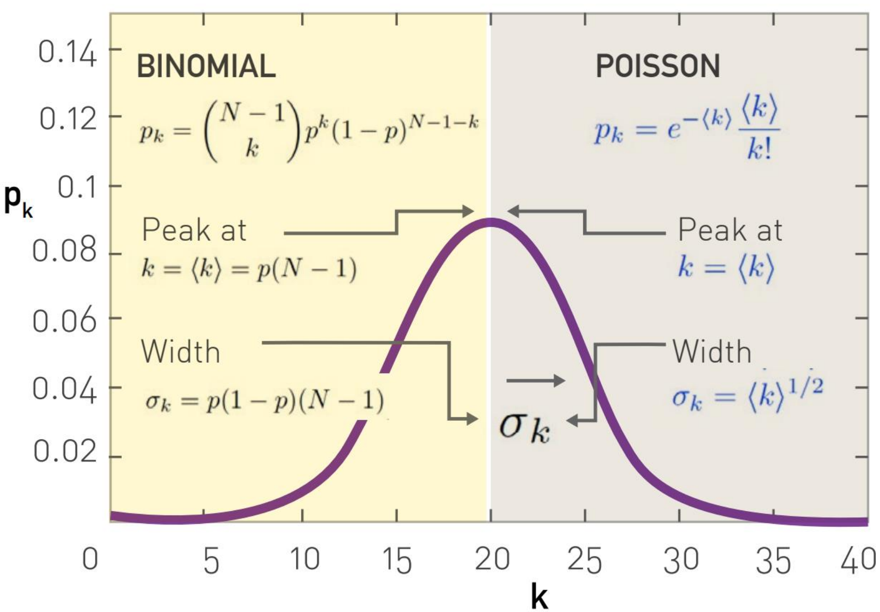
probability that the remaining (N-1-k) links are missing

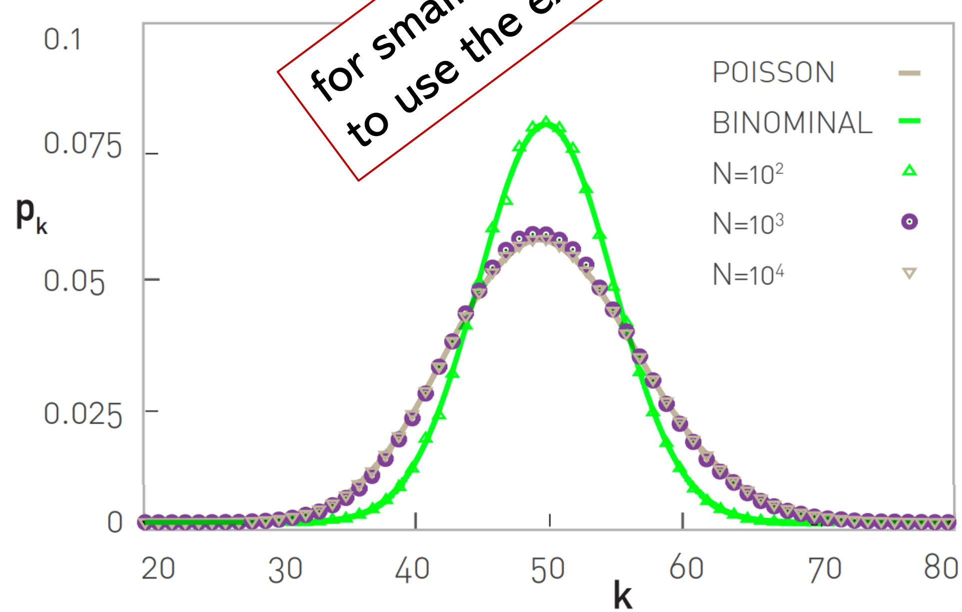
… however, we will instead use a Poisson distribution

- For most real networks $\langle k \rangle \ll N$
- In this limit, the degree distribution is well approximated by the Poisson distribution

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

for small networks one needs to use the exact binomial form





In a random network, can I have can high degree nodes and low degree nodes?

In a random network, the chance of observing a hub decreases faster than exponentially.

In a large random network the degree of most nodes is in the narrow vicinity of $\langle k \rangle$

Sociologists estimate that a typical person knows about 1,000 individuals on a first name basis, prompting us to assume that $<k> \approx 1{,}000$. Using the results obtained so far about random networks, we arrive to a number of intriguing conclusions about a random society of $N \simeq 7 \times 10^9$ of individuals (ADVANCED TOPICS 3.B):

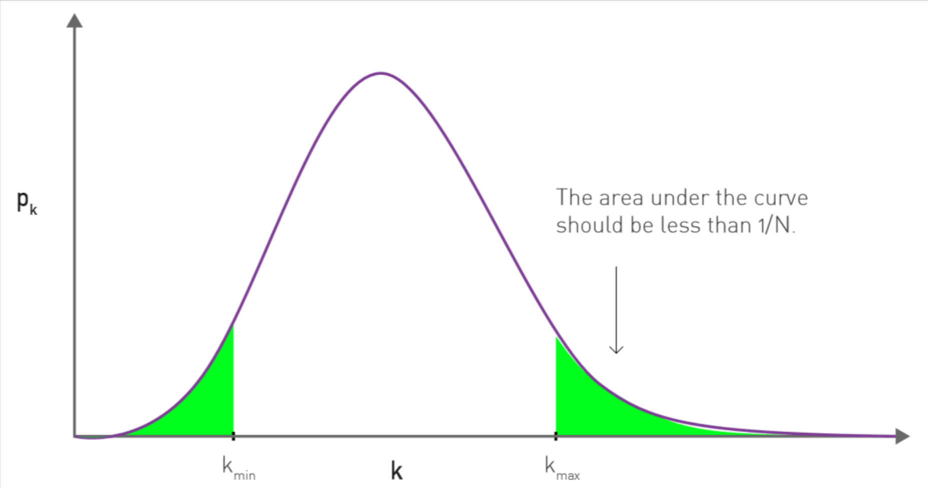
- The most connected individual (the largest degree node) in a random society is expected to have $k_{max} = 1{,}185$ acquaintances.
- The degree of the least connected individual is $k_{min} = 816$, not that different from k_{max} or $<k>$.
- The dispersion of a random network is $\sigma_k = <k>^{1/2}$, which for $<k> = 1{,}000$ is $\sigma_k = 31.62$. This means that the number of friends a typical individual has is in the $<k> \pm \sigma_k$ range, or between 968 and 1,032, a rather narrow window.

How to calculate network upper/lower natural cutoff

- network of N nodes
- k_{max} is the value of k such that we have at most one node with degree higher than k_{max}

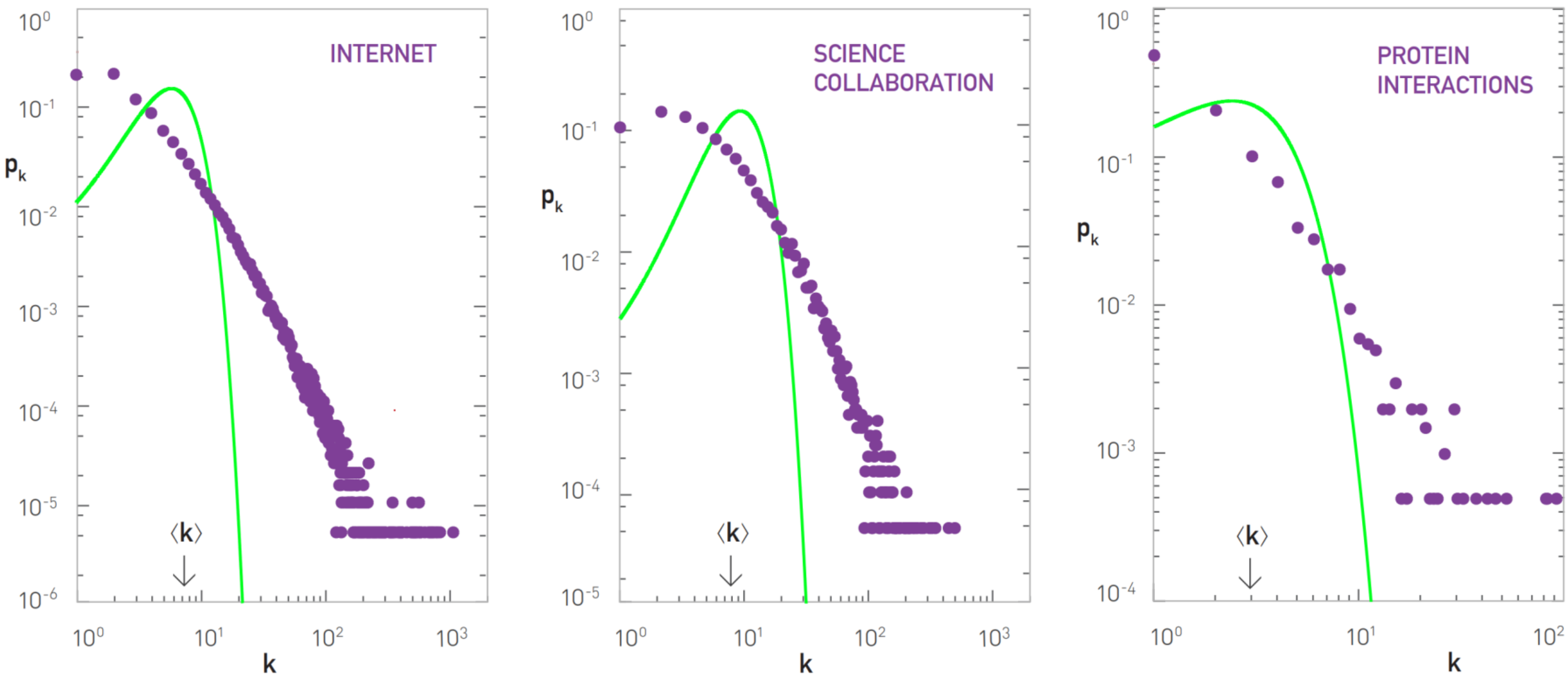
To calculate it, solve for k_{max} :

$$1 - P(k_{max}) \approx \frac{1}{N}$$



where $P(k)$ is the cumulative degree distribution.

Real networks are not Poisson



The random network model:

- underestimates the size and the frequency of the high degree nodes, as well as the number of low degree nodes
- predicts a larger number of nodes in the vicinity of $\langle k \rangle$ than seen in real networks

A.PACCANARO – FGV INTERNAL

The evolution of a Random Network

Networks evolve: Starting with N isolated nodes, links are added gradually. This corresponds to an increase in *p*.

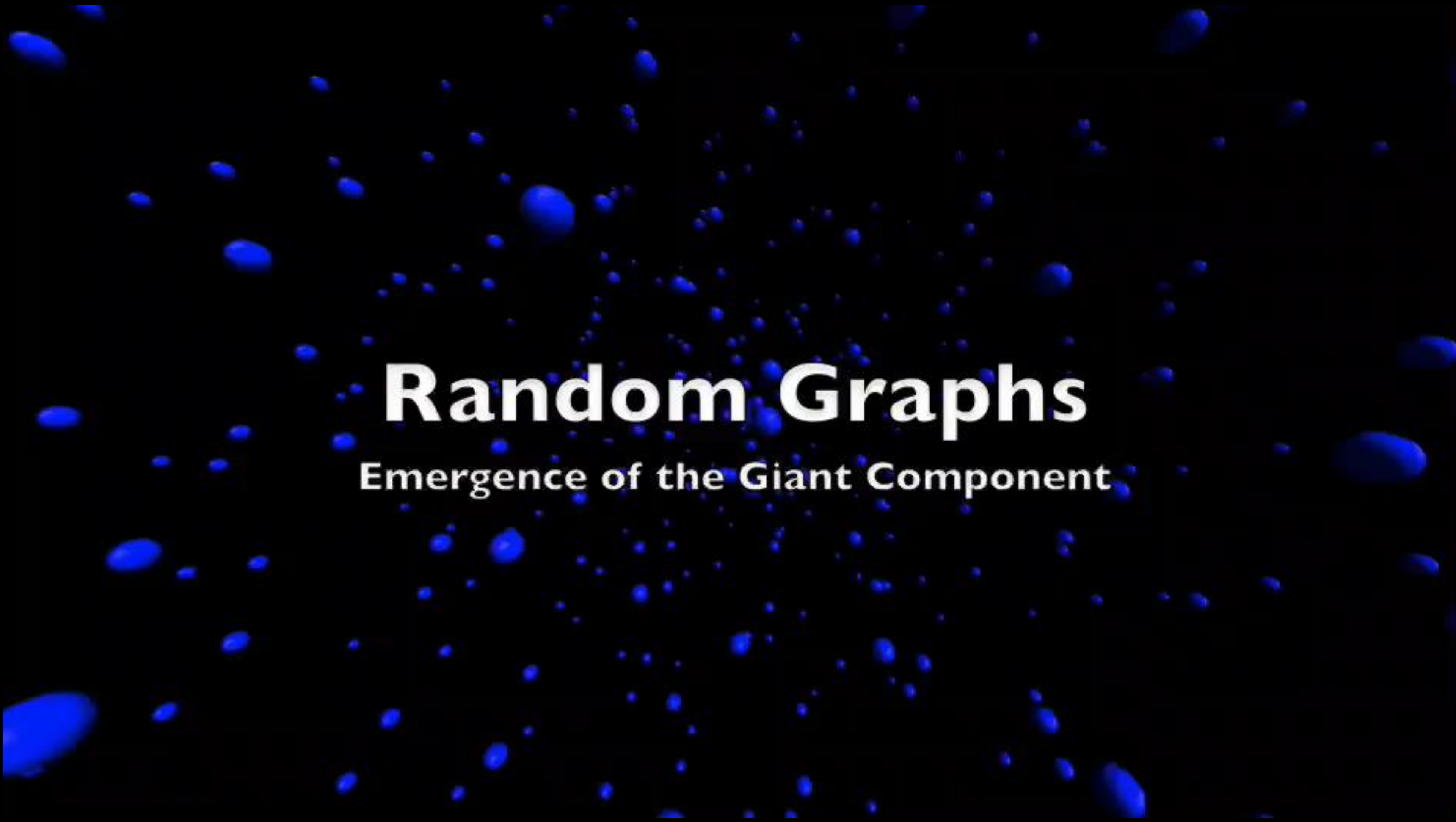
***Question**: how does the size of the largest connected cluster within the network, N_G, varies with $\langle k \rangle$?*

$$p = 0 \ \Rightarrow \langle k \rangle = 0 \text{ , all nodes are isolated, } \Rightarrow N_G = 1, \ \frac{N_G}{N} \longrightarrow 0 \text{ for large N}$$

$$p = 1 \ \Rightarrow \langle k \rangle = N - 1 \text{, all nodes are connected, } \Rightarrow N_G = N, \ \frac{N_G}{N} \longrightarrow 1$$

One would expect N_G to grow gradually from $N_G = 1$ to $N_G = N$ if $\langle k \rangle$ increases from 0 to N-1…

Instead $\frac{N_G}{N}$ grows suddenly, once $\langle k \rangle$ exceeds 1.



Random Graphs

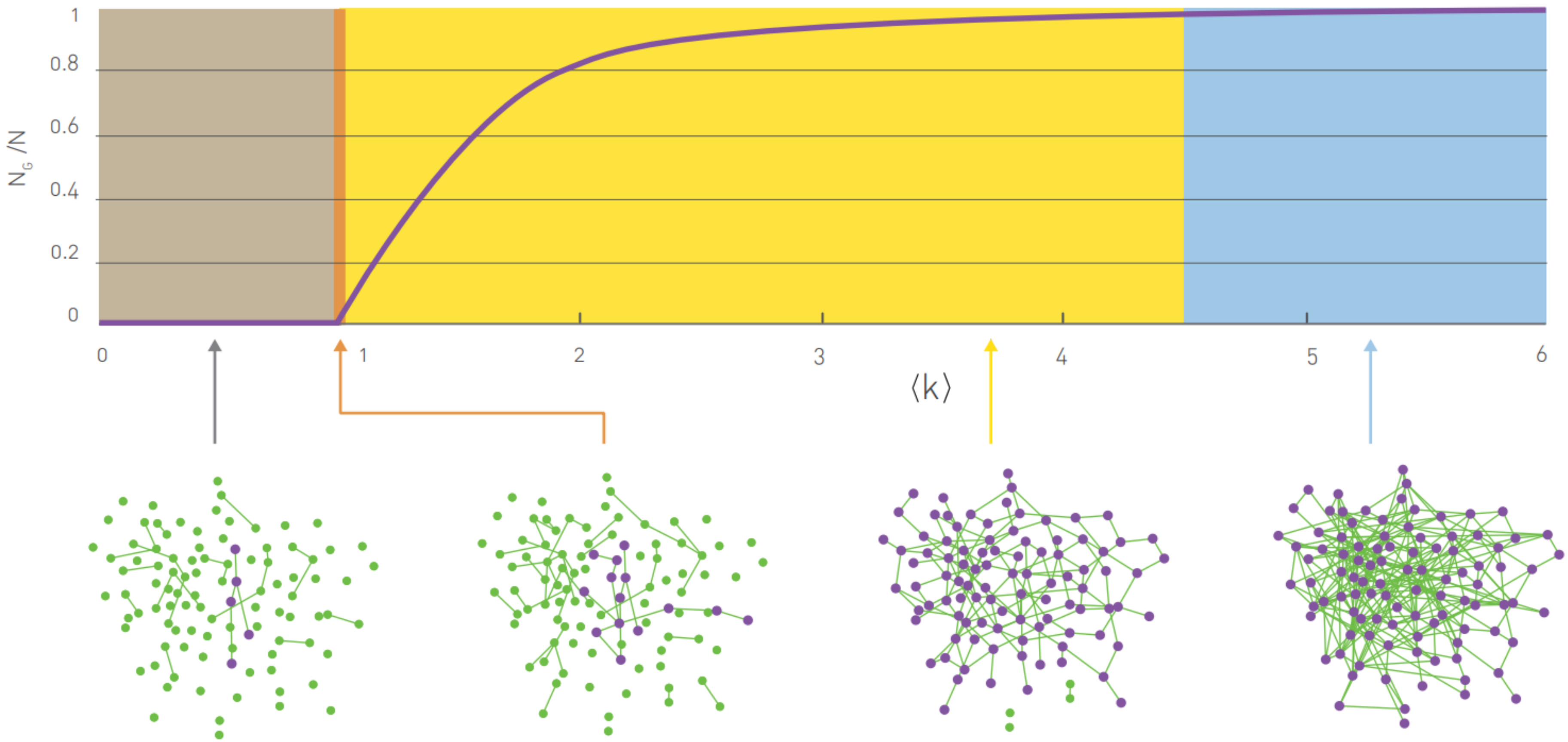
Emergence of the Giant Component

E&R result: we have a giant component if and only if each node has on average more than one link

$$\langle k \rangle = p(N-1) \implies p_c = \frac{1}{N-1} \approx 1/N$$

The larger a network, the smaller p is sufficient for the giant component.

- the emergence of a network is not a smooth, gradual process.
- 4 topologically distinct regimes



$\langle k \rangle < 1$ | $\langle k \rangle = 1$ | $\langle k \rangle > 1$ | $\langle k \rangle \gg \ln N$

(b) Subcritical Regime
- No giant component
- Cluster size distribution: $p_s \sim s^{-3/2} e^{-\alpha s}$
- Size of the largest cluster: $N_G \sim \ln N$
- The clusters are trees

Numerous tiny components with comparable sizes

(c) Critical Point
- No giant component
- Cluster size distribution: $p_s \sim s^{-3/2}$
- Size of the largest cluster: $N_G \sim N^{2/3}$
- The clusters may contain loops

Numerous small components with different sizes (mainly trees) one larger component with loops

(d) Supercritical Regime
- Single giant component
- Cluster size distribution: $p_s \sim s^{-3/2} e^{-\alpha s}$
- Size of the giant component: $N_G \sim (p - p_c)N$
- The small clusters are trees
- Giant component has loops

Numerous components coexist with the giant component. Small components are trees. Giant component has loops and cycles.

(e) Connected Regime
- Single giant component
- No isolated nodes or clusters
- Size of the giant component: $N_G = N$
- Giant component has loops

The network becomes connected. It is still relatively sparse.

A.P.

Are these predictions verified in real networks ?

- *Do real networks satisfy the criteria for the existence of a giant component, i.e.* $\langle k \rangle > 1$ *?*

NETWORK	N	L	$\langle k \rangle$	lnN
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	94,439	8.08	10.05
Actor Network	702,388	29,397,908	83.71	13.46
Protein Interactions	2,018	2,930	2.90	7.61

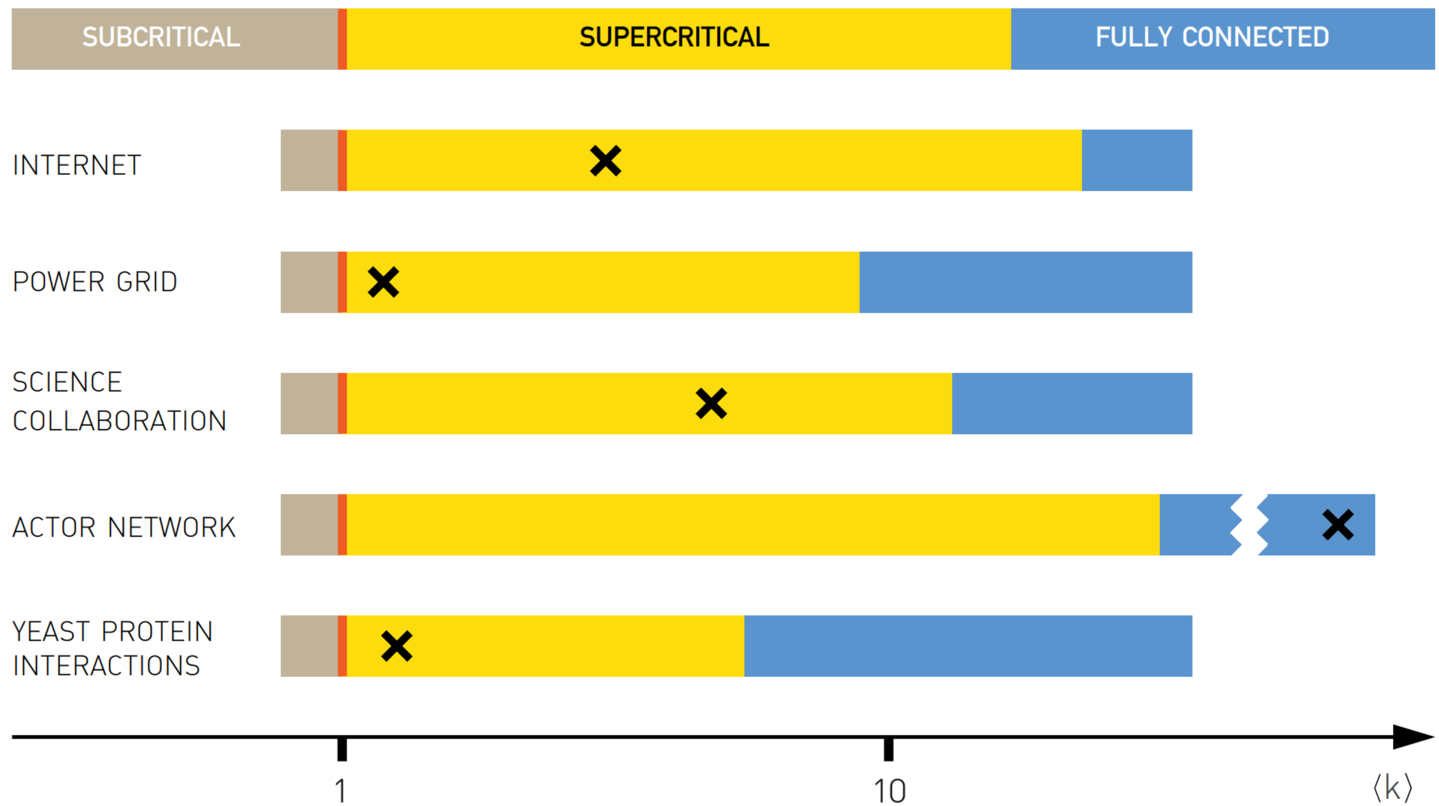
Yes, and they all have a giant component ✓

- *Do we see a single component (i.e. if $\langle k \rangle > \ln N$), or is the network fragmented into multiple components (i.e. if $\langle k \rangle < \ln N$).*

Many real networks do not obey the fully connected criteria $\langle k \rangle < \ln N$.

NETWORK	N	L	$\langle k \rangle$	lnN
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	94,439	8.08	10.05
Actor Network	702,388	29,397,908	83.71	13.46
Protein Interactions	2,018	2,930	2.90	7.61

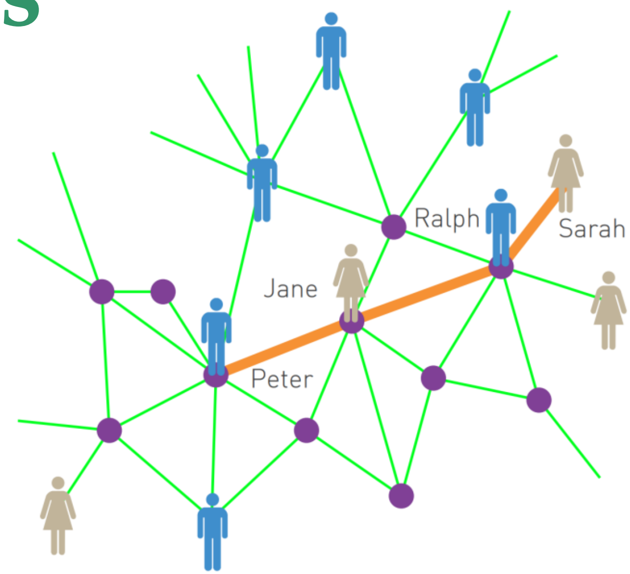
According to random network theory these networks should have a giant component BUT also be fragmented into several disconnected components



Most networks are in the supercritical regime, hence they are expected to be broken into numerous isolated components – and this does not happen in reality.

Small Worlds

Six degrees of separation: if you choose any two individuals anywhere on Earth, you will find a path of at most six acquaintances between them



Random network with average degree <*k*> :

<*k*> nodes at distance one (*d*=1).

<*k*>^{2} nodes at distance two (*d*=2).

<*k*>^{3} nodes at distance three (*d* =3).

...

<*k*>^{d} nodes at distance *d*.

$$N(d) \approx 1 + \langle k \rangle + \langle k \rangle^2 + \ldots + \langle k \rangle^d = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1}$$

expected number of nodes up to distance *d* from the starting node

N(d) cannot exceed *N*, hence we can identify the maximum distance in the network d_{max} :

$$N(d_{max}) \approx N \qquad \langle k \rangle^{d_{max}} \approx N$$

$$d_{max} \approx \frac{\ln N}{\ln \langle k \rangle}$$ diameter of a random network

Often, it is a better approximation to the average distance between two randomly chosen nodes, $\langle d \rangle$, than to d_{max}:

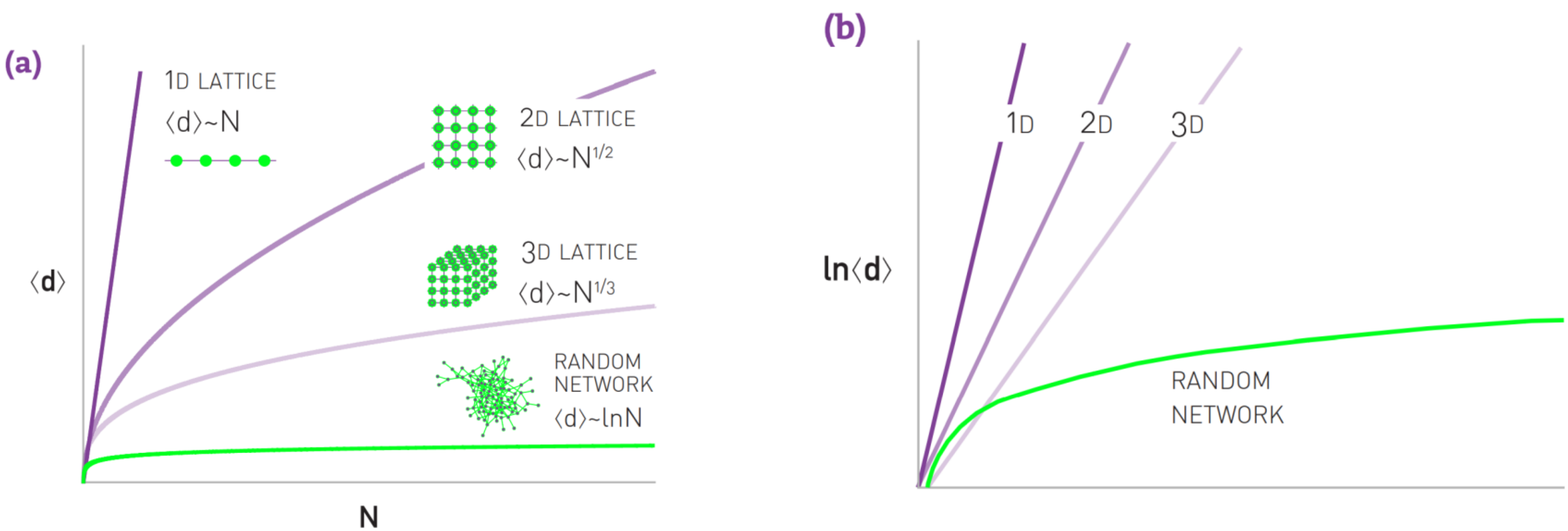
$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$$

small world property

NETWORK	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\ln N}{\ln \langle k \rangle}$
Internet	192,244	609,066	6.34	6.98	26	6.58
WWW	325,729	1,497,134	4.60	11.27	93	8.31
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.72	39	11.42
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	93,439	8.08	5.35	15	4.81
Actor Network	702,388	29,397,908	83,71	3,91	14	3,04
Citation Network	449,673	4,707,958	10.43	11,21	42	5.55
E. Coli Metabolism	1,039	5,802	5.58	2.98	8	4.04
Protein Interactions	2,018	2,930	2.90	5.61	14	7.14

- distances in a random network are orders of magnitude smaller than the size of the network
- denser the network, the smaller is the distance between the nodes.

$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$$



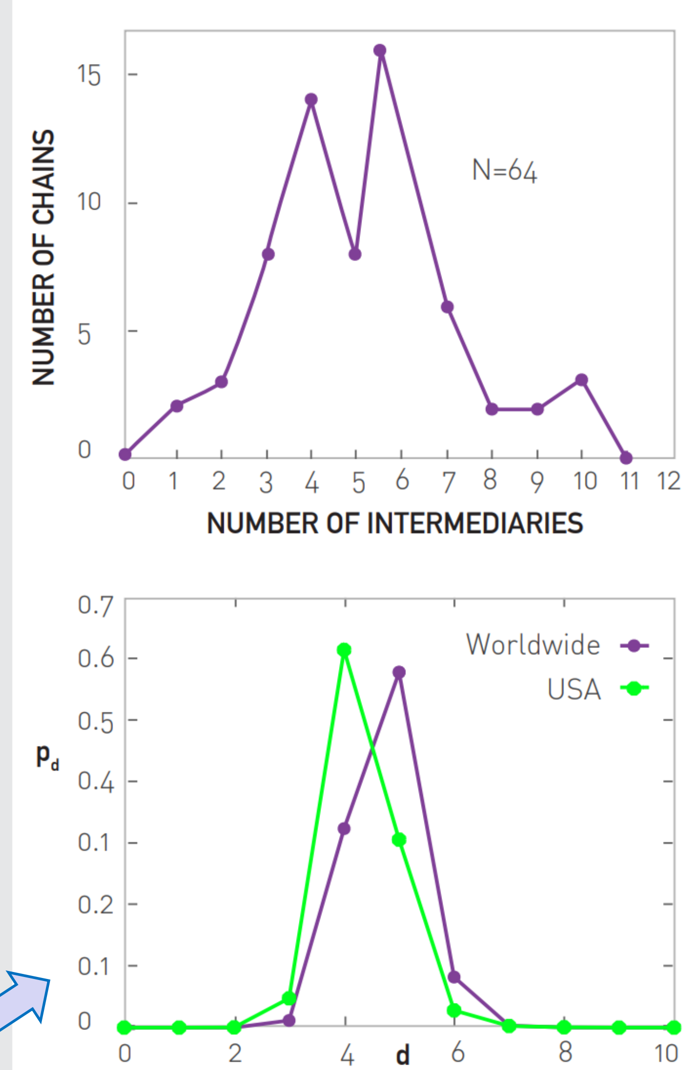
For social networks, using $N \approx 7 \times 10^9$ and $\langle k \rangle \approx 10^3$ ➔ *3.28* !!!

The small world phenomena -- rooted in the fact that the number of nodes at distance d from a node increases exponentially with d.

SIX DEGREES: EXPERIMENTAL CONFIRMATION

The first empirical study of the small world phenomena took place in 1967, when Stanley Milgram, building on the work of Pool and Kochen [20], designed an experiment to measure the distances in social networks [24, 25]. Milgram chose a stock broker in Boston and a divinity student in Sharon, Massachusetts as *targets*. He then randomly selected residents of Wichita and Omaha, sending them a letter containing a short summary of the study's purpose, a photograph, the name, address and information about the target person. They were asked to forward the letter to a friend, relative or acquantance who is most likely to know the target person.

Within a few days the first letter arrived, passing through only two links. Eventually 64 of the 296 letters made it back, some, however, requiring close to a dozen intermediates [25]. These completed chains allowed Milgram to determine the number of individuals required to get the letter to the target (Figure 3.12a). He found that the median number of intermediates was 5.2, a relatively small number that was remarkably close to Frigyes Karinthy's 1929 insight (BOX 3.8).



distance distribution, p_d , for pairs of Facebook users (2011)

Clustering coefficient of a RN

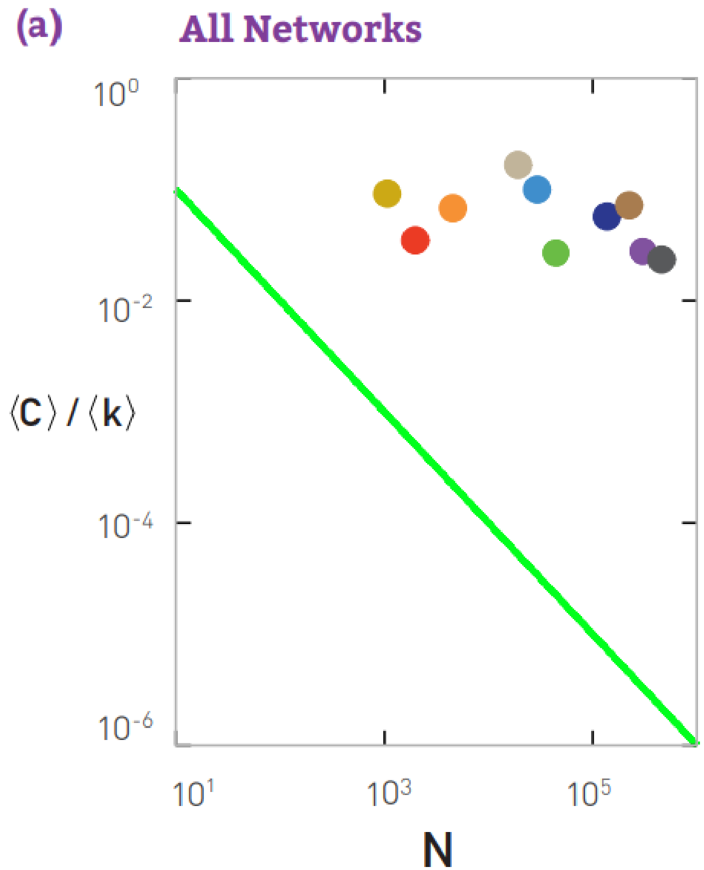
Expected number of links L_i between the node's k_i neighbours:

$$\langle L_i \rangle = p \frac{k_i(k_i - 1)}{2}$$

$$\Longrightarrow \qquad C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}$$

Prediction of RN model: For fixed <*k*>, the larger the network, the smaller is a node's clustering coefficient.

Plot $\langle C \rangle / \langle k \rangle$ in function of N



$\langle C \rangle / \langle k \rangle$ does not decrease as N^{-1}

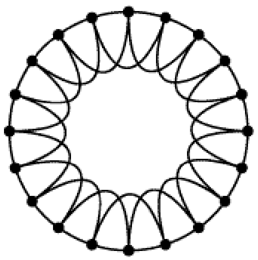
it is largely independent of N

real networks have a much higher clustering coefficient than expected for a random network

The random network model does not capture the clustering of real network

Watts & Strogatz: the idea/the question

Regular



?

Collective dynamics of 'small-world' networks

Duncan J. Watts* & Steven H. Strogatz

Department of Theoretical and Applied Mechanics, Kimball Hall, Cornell University, Ithaca, New York 14853, USA

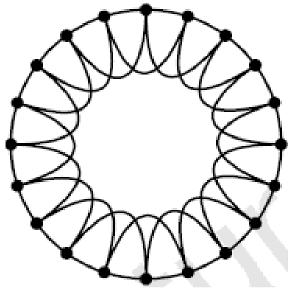
Nature, Vol. 393, 440, 1998

REWIRING PROCEDURE

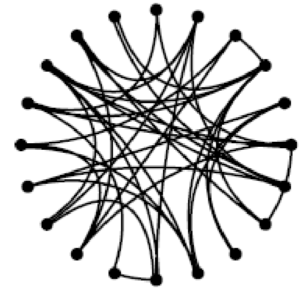
- Start with a regular network with N vertices
- Rewire each edge with probability p

p=0 ➔ regularity
p=1 ➔ disorder (random)
Question: what happens for 0<p<1 ?

Quantify the structural properties of the graph by its characteristic path length *L(p)* and clustering coefficient *C(p)*



n vertices
k edges per vertex



For p → 0 (Regular Networks):

- $L \sim n/2k >> 1$
- $C \sim ½$ (here)

- high clustering coefficient
- high characteristic path length

- highly clustered
- large world [L grows lin. with n]

For p → 1:

- $L \sim \ln(n)/\ln(k)$
- $C \sim k/n << 1$

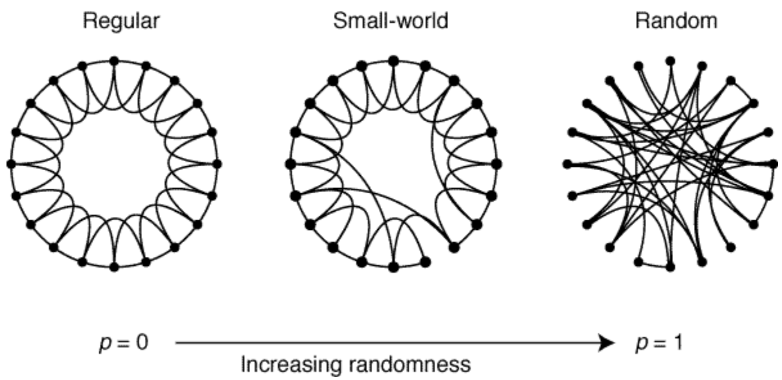
- low clustering coefficient
- low characteristic path length

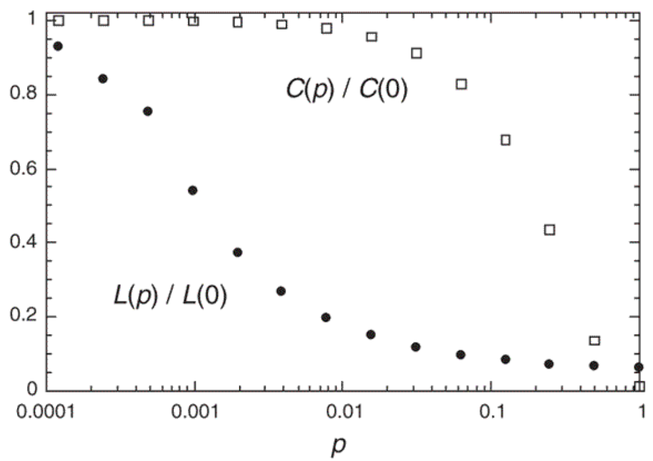
- poorly clustered
- small world [L grows log. with n]

This might lead to believe that large C is always associated with large L, and small C with small L…

A.PACCANARO

(1) There is a broad interval of p for which L is small but C remains large





(2) Hypothesis: small-world property might be common in sparse networks with many vertices as even a tiny fraction of short cuts could be sufficient

Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

Comparison with random graphs with the same number of vertices n and average degree k

Actors:	n=225226	k=61
Power grid:	n=4941	k=2.67
C.Elegans:	n=282	k=14

Conclusions

- The small-world phenomenon is not merely a curiosity of social networks nor an artefact of an idealized model --- ***it is probably generic for many large, sparse networks found in nature***

- The distinctive combination of high clustering with short characteristic path length in small-world networks **cannot be captured by traditional approximations** such as those based on regular lattices or random graphs.

(From Watts and Strogatz, 1998)