Random Networks

Vander Luis de Souza Freitas
vander.freitas@ufop.edu.br

UFOP

Agenda

- Last class we discussed:
  - Graph theory
  - Degree
  - General network characteristics
  - Paths and distances
  - Clustering coefficient
  - Coding!

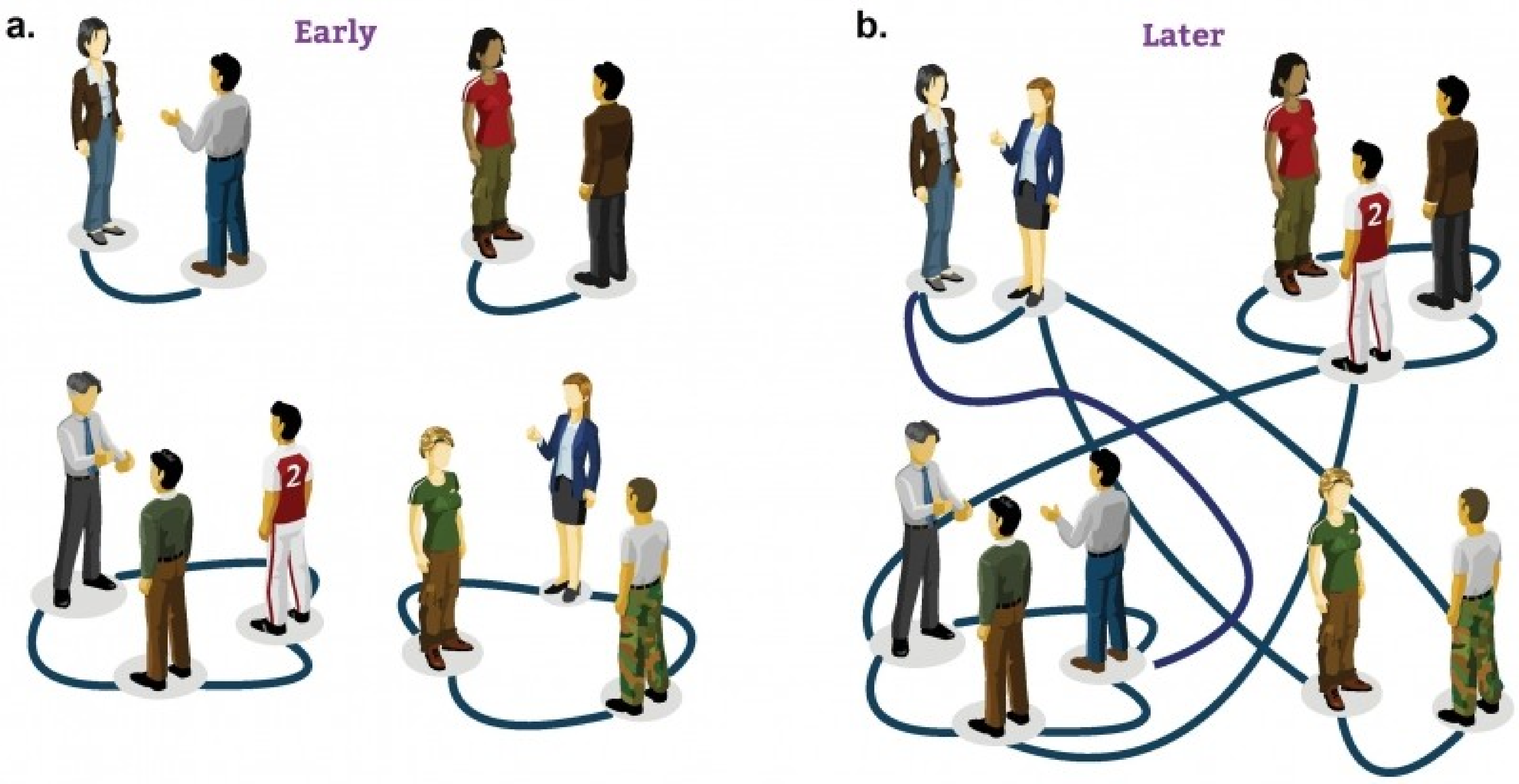
Agenda

- Today:
  - Discussion about the homework
  - Random networks

Last class homework

Discussion

From a Cocktail Party to Random Networks



http://networksciencebook.com/chapter/3

Random networks

***G(N,L)* model**: *N* labeled nodes are connected with *L* randomly placed links (ERDÖS; RÉNYI, 1959).

G(N,p) model: each pair of *N* labeled nodes is connected with probability *p* (GILBERT, 1959).

Regardless of the way one builds the networks, people commonly call random networks: Erdös-Rényi network, or simply ER network.

P. Erdős and A. Rényi. On random graphs, I. Publicationes Mathematicae (Debrecen), 6:290-297, 1959.

E. N. Gilbert. Random graphs. The Annals of Mathematical Statistics, 30:1141-1144, 1959.

http://networksciencebook.com/chapter/3

Random networks

G(N,p) model: each pair of *N* labeled nodes is connected with probability *p* (GILBERT, 1959).

Algorithm:
1) Start with *N* isolated nodes.
2) Select a node pair and generate a random number between *0* and *1*. If the number is less than *p*, connect the selected node pair with a link, otherwise leave them disconnected.
3) Repeat step (2) on each of the *N(N-1)/2 possible* node pairs.

http://networksciencebook.com/chapter/3

Number of links

Parameters *N* and *p* generate different networks for each realization and often with different number of links *L*.

Probability that *L* of the attempts to connect the *N(N-1)/2* pairs of nodes have resulted in a link (each): p^L

Probability that the remaining *N(N-1)/2 - L* attempts have not resulted in a link: $(1-p)^{N(N-1)/2-L}$

Probability that a particular realization of a random network has exactly *L* links:

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

Recall that:

$$\binom{N}{x} = \mathrm{N!}\frac{}{x!(N-x)!}$$

http://networksciencebook.com/chapter/3

Binomial distribution

Each realization has two possible outcomes with probabilities *p* and *1-p* respectively. Using the tosses of a coin as an example, the probability that we obtain exactly *x* heads in sequence of *N* throws is:

$$p_x = \binom{N}{x} p^x (1-p)^{N-x}$$

The mean of the distribution (first moment):

$$\langle x \rangle = \sum_{x=0}^{N} x p_x = Np$$

Binomial distribution

Its second moment is:

$$\langle x^2 \rangle = \sum_{x=0}^{N} x^2 p_x = p(1-p)N + p^2 N^2$$

providing its standard deviation as:

$$\sigma_x = \left(\langle x^2 \rangle - \langle x \rangle^2 \right)^{\frac{1}{2}} = [p(1-p)N]^{\frac{1}{2}}$$

http://networksciencebook.com/chapter/3

Number of links

As p_L is a binomial distribution, the expected number of links is:

$$\langle L \rangle = \sum_{L=0}^{\frac{N(N-1)}{2}} L p_L = p\frac{N(N-1)}{2}$$

$$p_x = \binom{N}{x} p^x (1-p)^{N-x}$$

$$\langle x \rangle = \sum_{x=0}^{N} x p_x = Np$$

with

$$L_{max} = N(N-1)/2$$

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

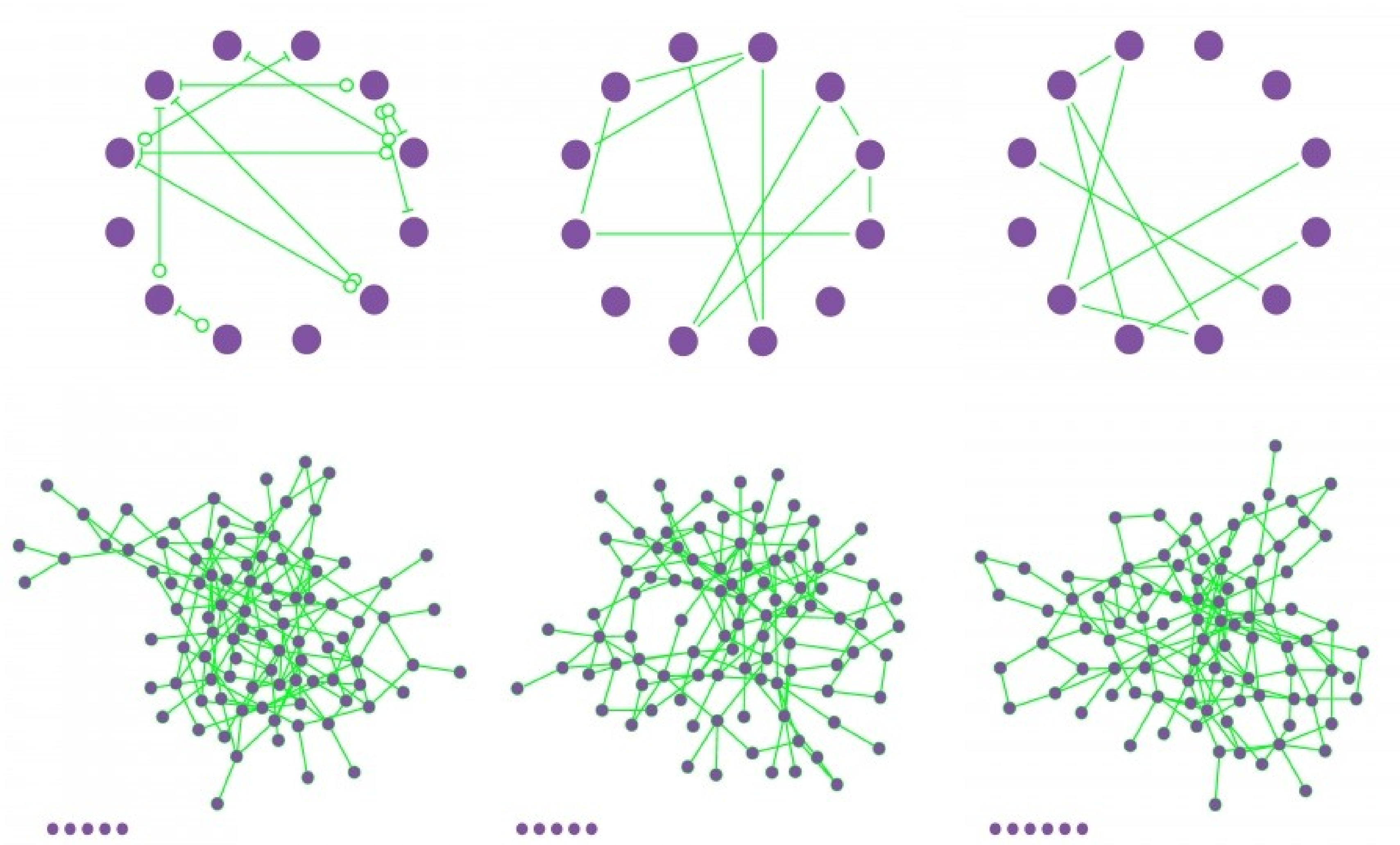
The average degree is therefore:

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

http://networksciencebook.com/chapter/3

Number of links

Top: three realizations with *p=1/6* and *N=12*: (*L=10,10,8*).
Bottom: three realizations with *p=0.03* and *N=100*.



http://networksciencebook.com/chapter/3

Degree distribution

The degree distribution of a random network follows the binomial distribution:

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

Most real networks are sparse: $\langle k \rangle \ll N$

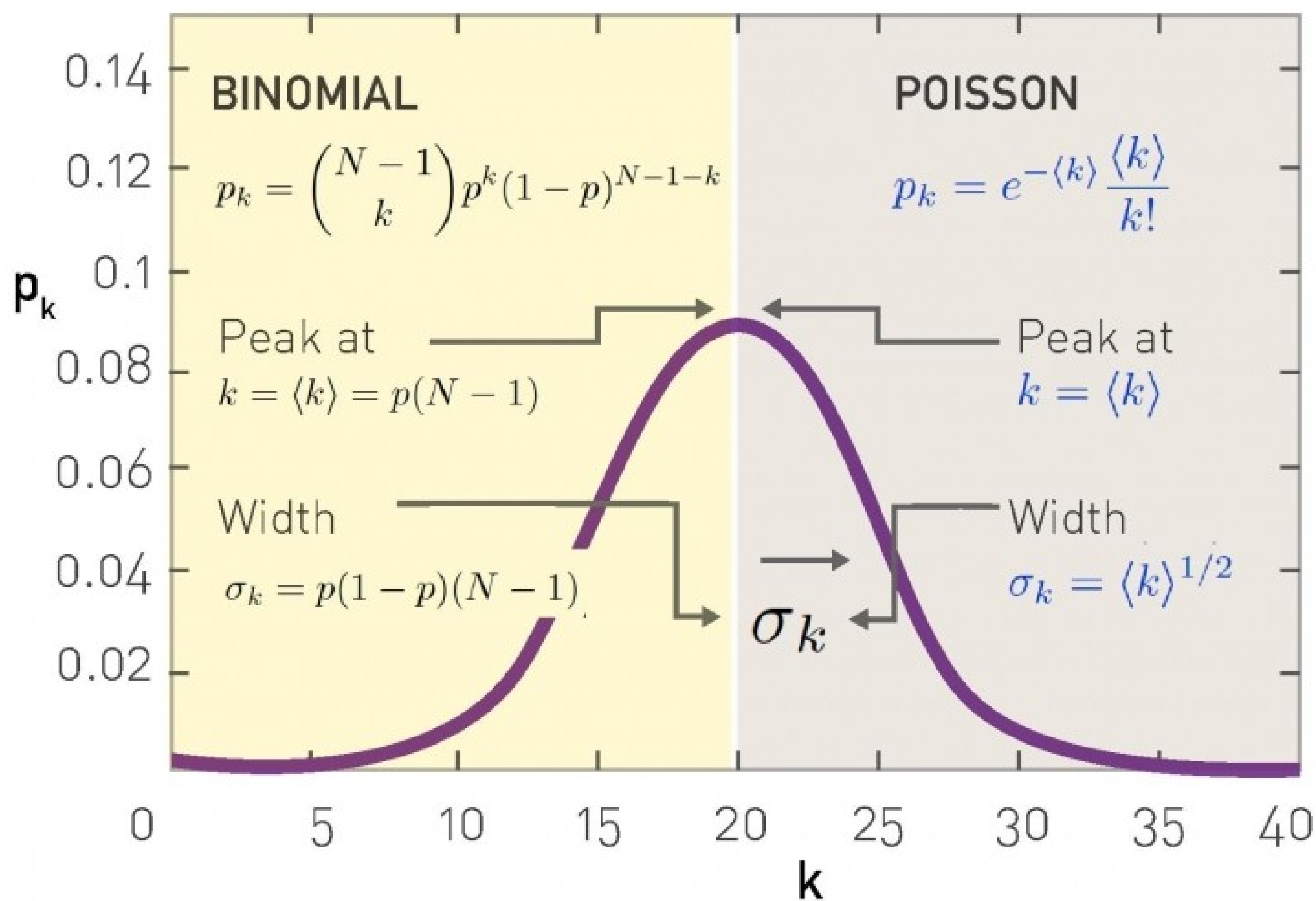
In this limit, p_k is well approximated by the Poisson distribution:

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

http://networksciencebook.com/chapter/3

Degree distribution

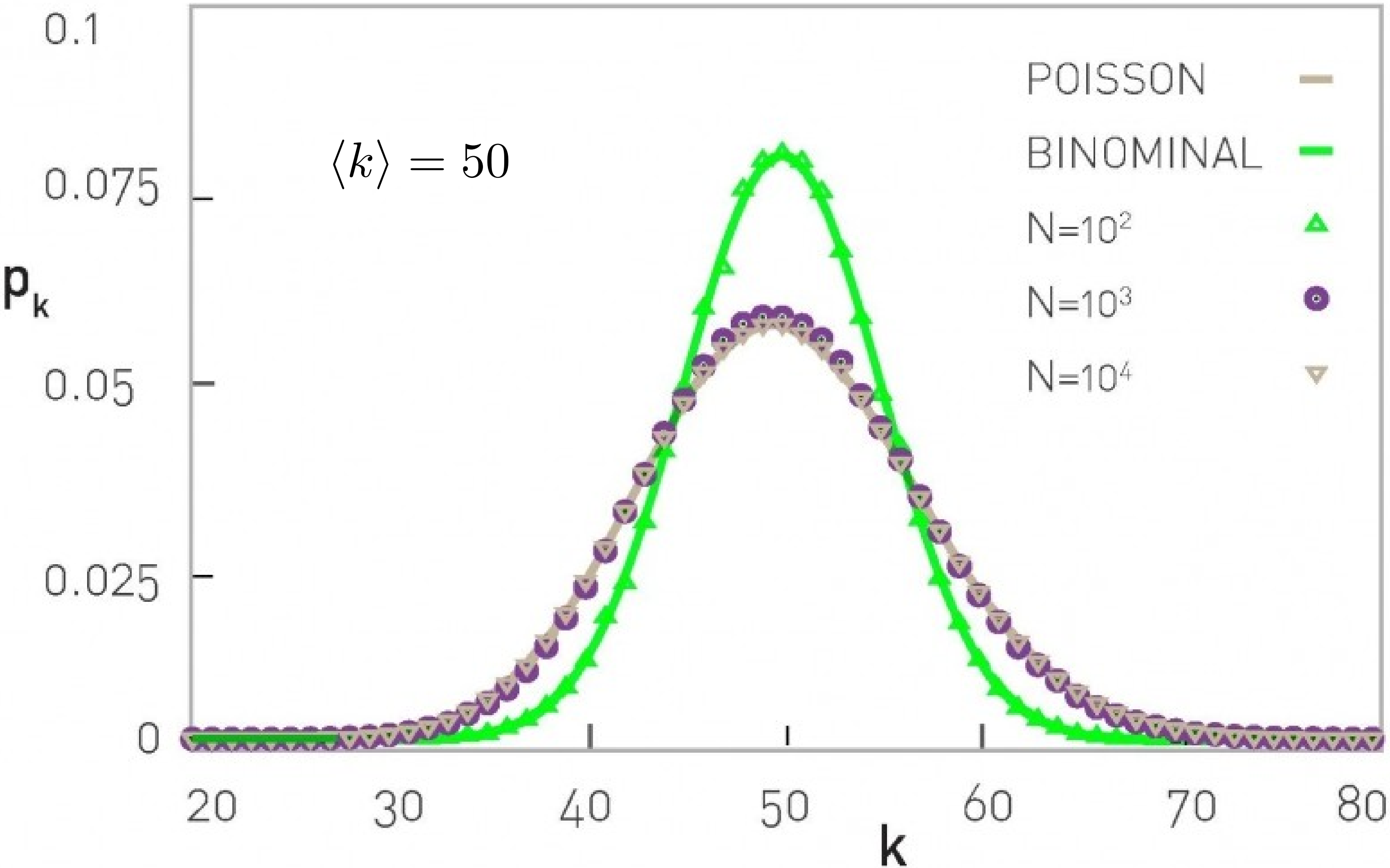
The exact result for the degree distribution is the binomial form, thus Poisson represents only an approximation that is valid for $\langle k \rangle \ll N$.



http://networksciencebook.com/chapter/3

Degree distribution

The Poisson distribution is commonly referred to as the "distribution of a random network". The main benefits are the simplicity of computing its statistics and the fact of depending on a single parameter.



http://networksciencebook.com/chapter/3

Real Networks are Not Poisson

Can high degree nodes coexist with small degree nodes?

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

Example:

- Assume we are modeling the world's social network with $N = 7 \times 10^9$.
- Sociologists estimate that a typical person knows about 1000 individuals on a first name basis, therefore $\langle k \rangle = 1000$.
- The dispersion is $\sigma_k = \langle k \rangle^{1/2}$ which gives $\sigma_k = 31.62$.
- This means a typical individual has about $\langle k \rangle \pm \sigma_k$ friends, that is between 968 and 1032.
- Not true for real networks! There are people with many more friends!

http://networksciencebook.com/chapter/3

Real Networks are Not Poisson

Remember the Poisson distribution

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

Note this factorial term

The Stirling approximation

$$k! \sim [\sqrt{2\pi k}] \left(\frac{k}{e}\right)^k$$

Rewriting p_k :

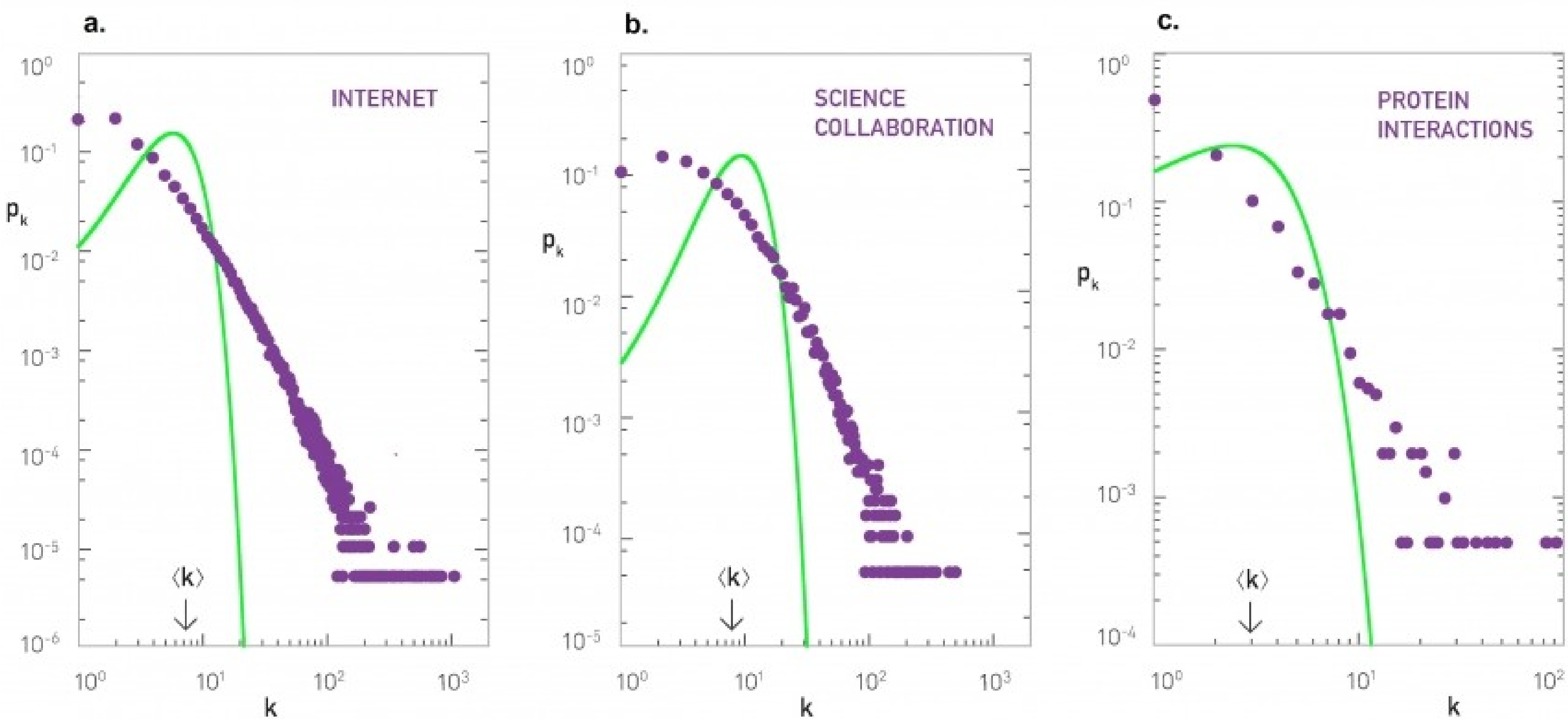
$$p_k = \frac{e^{-\langle k \rangle}}{\sqrt{2\pi k}} \left(\frac{e \langle k \rangle}{k}\right)^k$$

When $k > \langle k \rangle$, p_k
decreases considerably!

http://networksciencebook.com/chapter/3

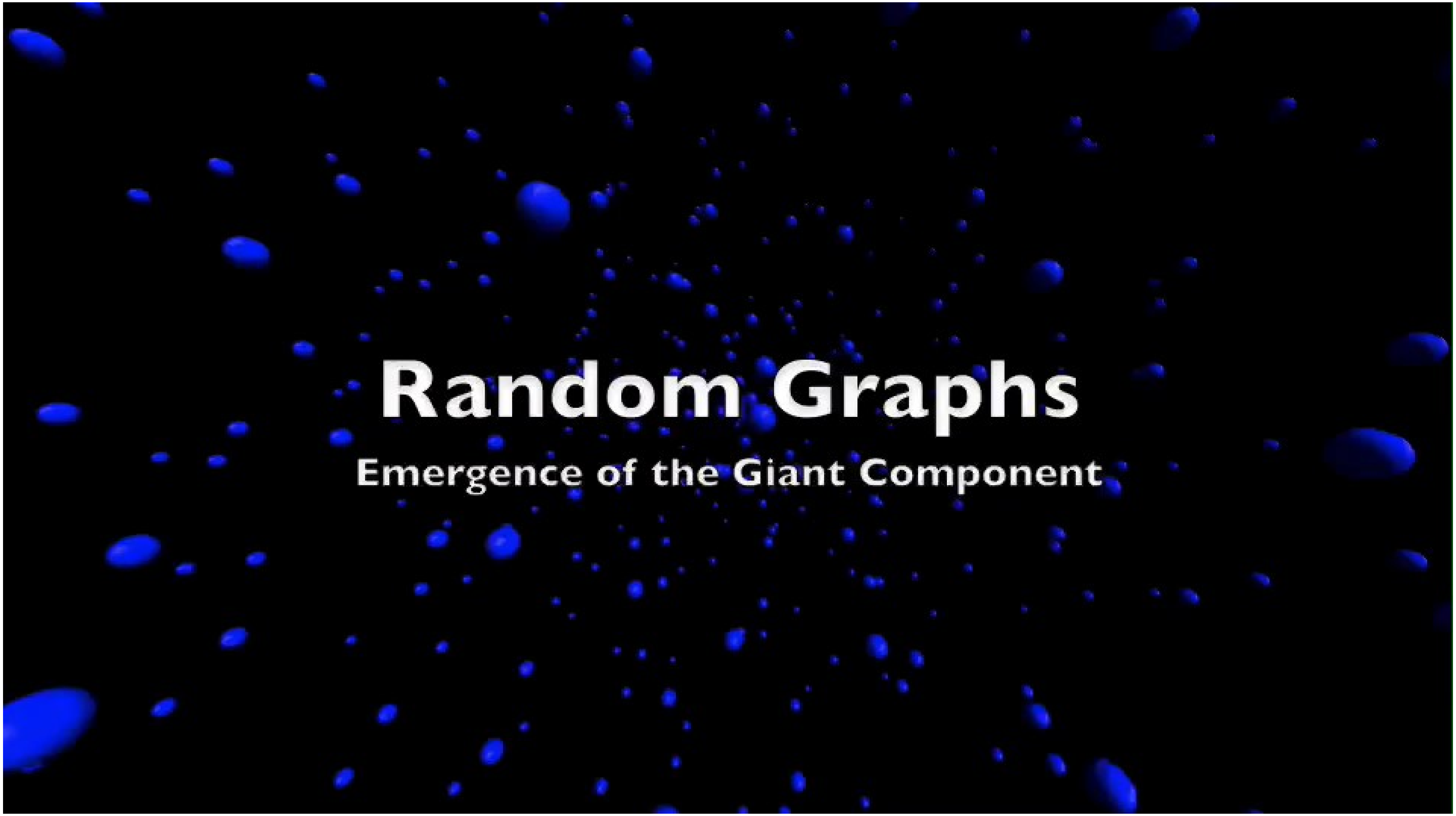
Degree distribution

Comparing the degree distribution of real networks with random networks:



http://networksciencebook.com/chapter/3

The evolution of a random network



http://networksciencebook.com/chapter/3

The evolution of a random network

When $p = 0 \rightarrow \langle k \rangle = 0$

this leads the size of the giant component to $N_G = 1$.

When $p = 1 \rightarrow \langle k \rangle = N - 1$, therefore $N_G = N$ and $N_G/N = 1$.

However N_G does not grow gradually. A giant component emerge at $\langle k \rangle = 1$, which gives the critical probability

$$p_c = \frac{1}{N-1} \approx \frac{1}{N}$$

Recall that:

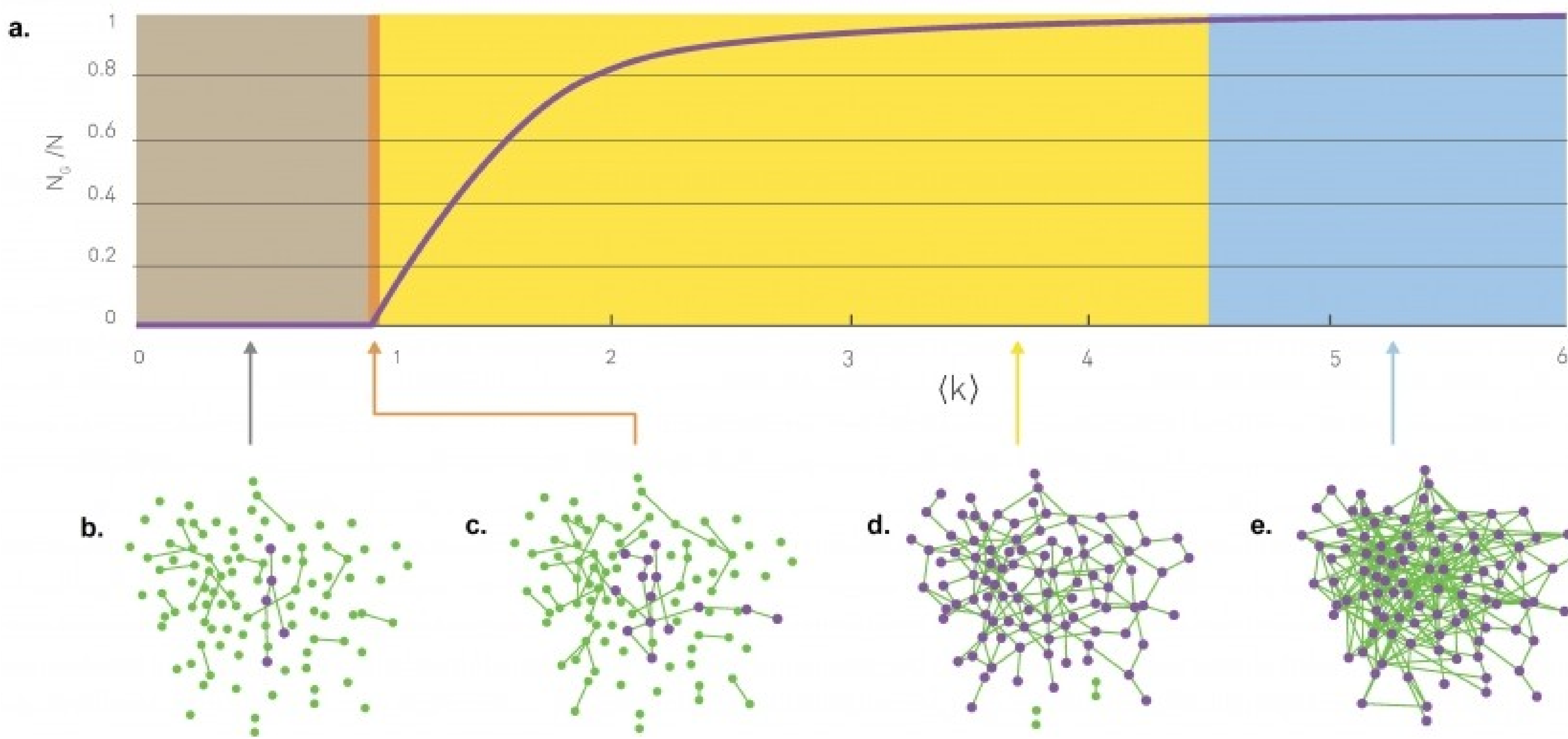
$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1)$$

Therefore the larger the network, the smaller *p* is sufficient for the giant component.

The evolution of a random network

Transitions:

Subcritical Regime | Critical Point | Supercritical Regime | Connected Regime



http://networksciencebook.com/chapter/3

The evolution of a random network

Subcritical regime: $0 < \langle k \rangle < 1 \quad \left(p < \frac{1}{N} \right)$

The size of the giant component increases much slower than the size of network:

$$N_G \sim \ln N$$

$N_G/N \simeq \ln N/N \rightarrow 0$ in the $N \rightarrow \infty$ limit.

This regime consists of numerous tiny components of comparable sizes.

The evolution of a random network

Critical Point: $\langle k \rangle = 1 \quad \left(p = \frac{1}{N}\right)$

Still growing slower than the network: $N_G \sim N^{2/3}$.

Its relative size decreases as $N_G/N \sim N^{-1/3}$ in the $N \to \infty$ limit.

Comparing with the Subcritical Regime:

- Example with $N = 7 \times 10^9$:
  - In the Subcritical Regime: $N_G \simeq 22.7$
  - At the Critical Point: $N_G \simeq 3 \times 10^6$

  - There is a jump in the giant component size and there are numerous small components (mainly trees) whose size distribution follows a power law.

http://networksciencebook.com/chapter/3

The evolution of a random network

Supercritical Regime: $\langle k \rangle > 1 \quad \left(p > \frac{1}{N}\right)$

Now there is a real giant component. In <u>the vicinity</u> of $\langle k \rangle = 1$

$$N_G / N \sim \langle k \rangle - 1$$

or

$$N_G \sim (p - p_c) N$$

where

$$p_c = \frac{1}{N-1} \approx \frac{1}{N}$$

For large $\langle k \rangle$ the dependence between N_G and $\langle k \rangle$ is nonlinear.
In this regime, the small components are usually trees and the giant component contains loops and cycles. It lasts until all nodes are absorbed by the giant component.

http://networksciencebook.com/chapter/3

The evolution of a random network

Connected Regime: $\langle k \rangle > \ln N \quad \left(p > \frac{\ln N}{N} \right)$

All nodes are within the only component, hence $N_G \simeq N$.
When we enter this regime, the network is still sparse. In only becomes a complete graph at $\langle k \rangle = N - 1$

Real networks are Supercritical

Supercritical Regime: $\langle k \rangle > 1$

Connected Regime: $\langle k \rangle > \ln N$

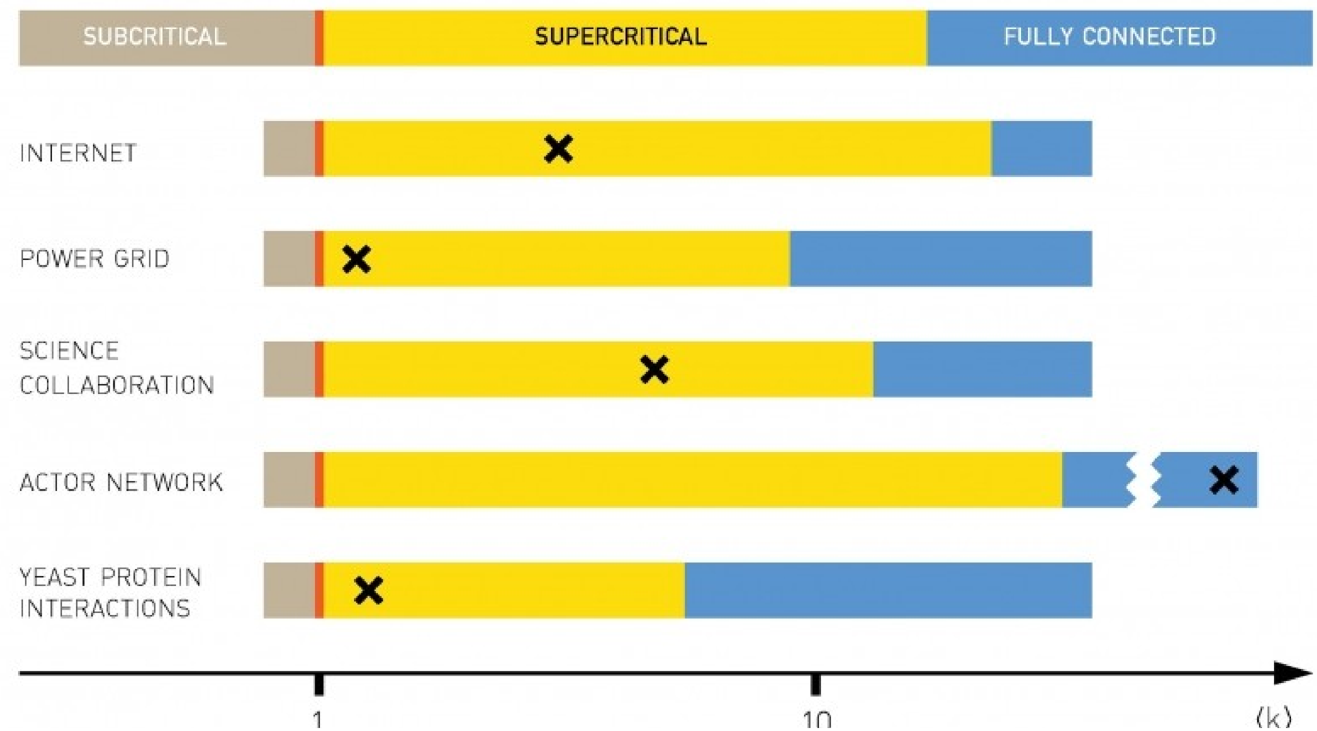
Network	N	L	‹K›	lnN
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	94,437	8.08	10.05
Actor Network	702,388	29,397,908	83.71	13.46
Protein Interactions	2,018	2,930	2.90	7.61

http://networksciencebook.com/chapter/3

Real networks are Supercritical

Supercritical Regime: $\langle k \rangle > 1$

Connected Regime: $\langle k \rangle > \ln N$



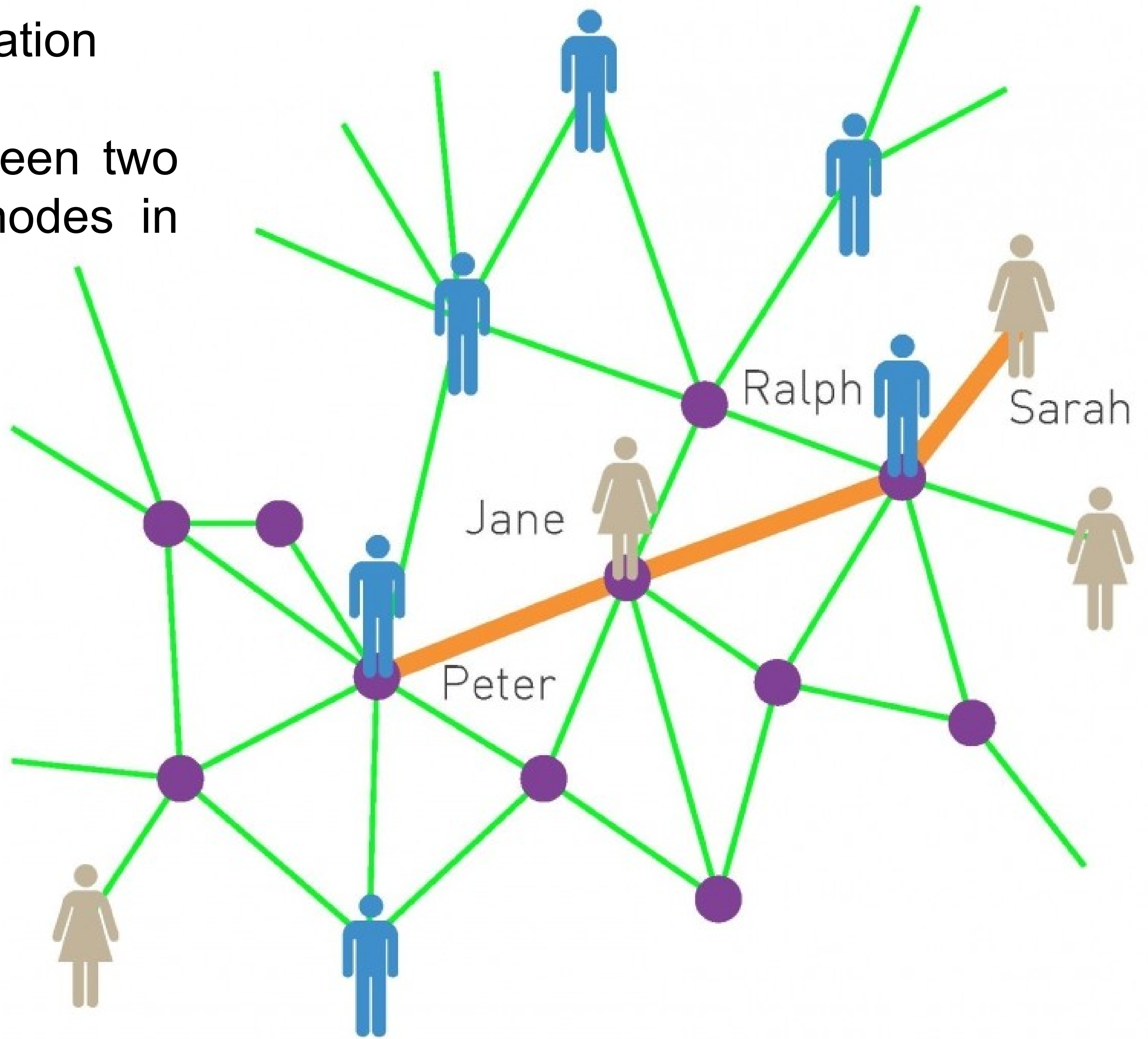
Question: Does this mean that real networks have giant components but are disconnected? This prediction fails for several real networks, because they usually do not agree with the ER model.

http://networksciencebook.com/chapter/3

Small Worlds

Six degree of separation

“The distance between two randomly chosen nodes in a network is short”.



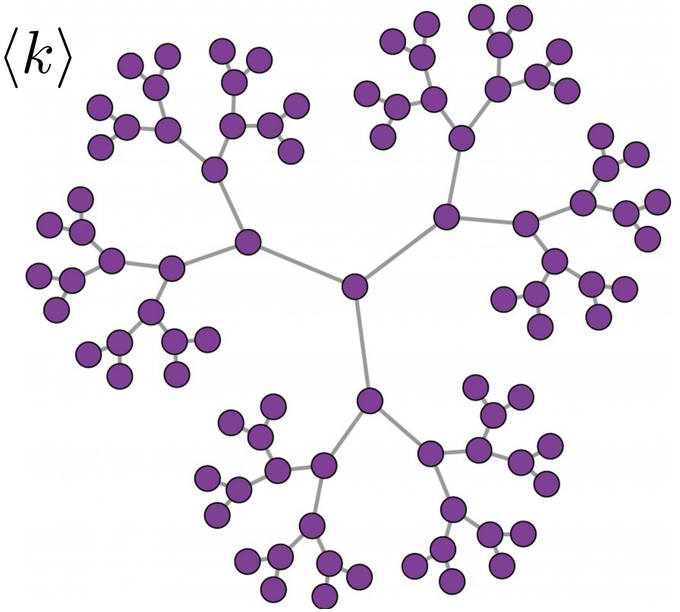
http://networksciencebook.com/chapter/3

Small Worlds

Consider a random network with average degree $\langle k \rangle$

A node in this network has on average:

- $\langle k \rangle$ nodes at distance one (*d=1*).
- $\langle k \rangle^2$ nodes at distance two (*d=2*).
- $\langle k \rangle^3$ nodes at distance three (*d=3*).
- …
- $\langle k \rangle^d$ nodes at distance *d*.



The expected nodes up to distance *d* from a starting node is

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

with $N(d) \leq N$. Then $N(d_{max}) \approx N$. Assuming $\langle k \rangle \gg 1$:

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

$$d_{max} \approx \frac{\ln N}{\ln \langle k \rangle}$$

The mathematical formulation of the small world phenomenon

http://networksciencebook.com/chapter/3

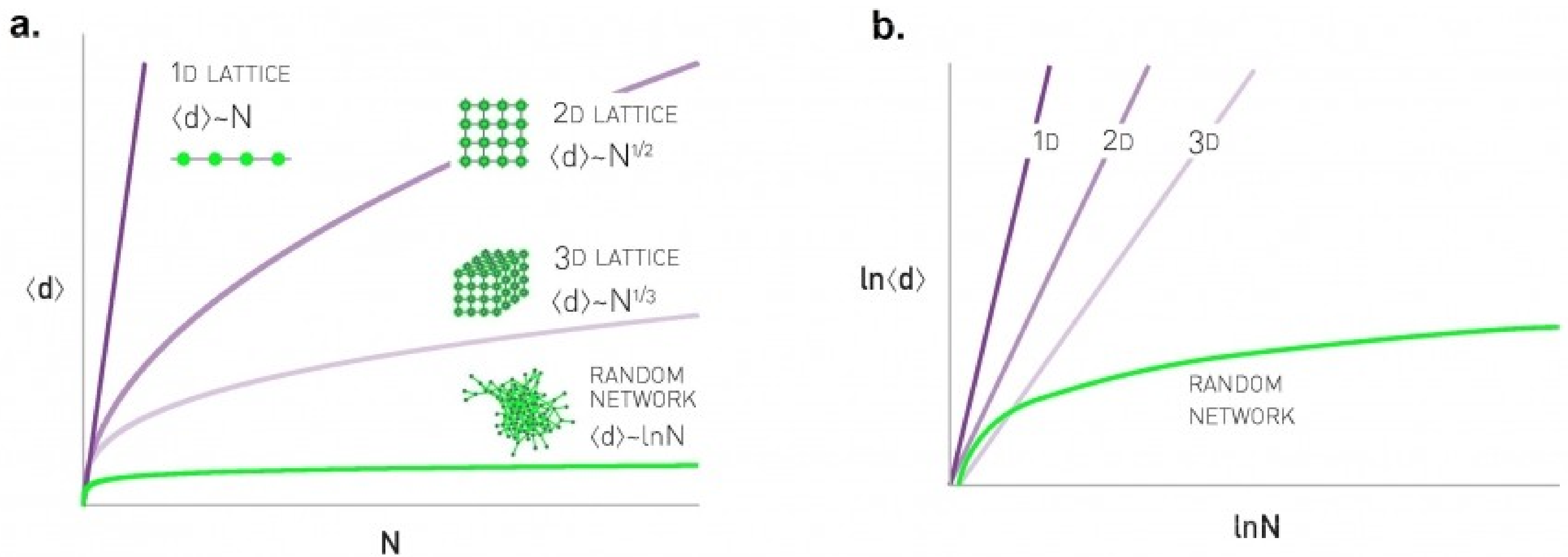
Small Worlds

However, we usually use

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

It means that the distances in a random network are “orders of magnitude smaller than the size of the network”.

“Small” means that the average path length or the diameter depends logarithmically on the system size. They are proportional to *ln N* rather than *N*.



http://networksciencebook.com/chapter/3

Small Worlds

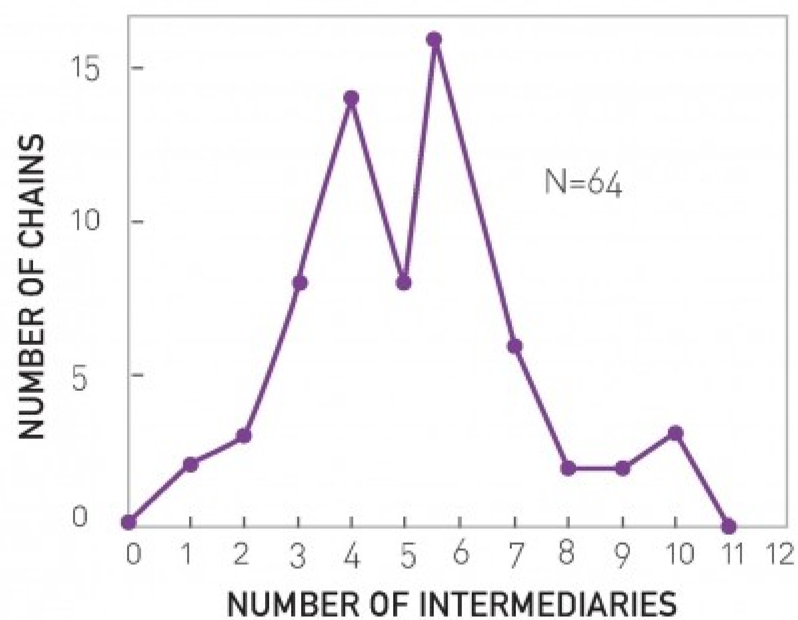
The last column gives the $\langle d \rangle \approx \dfrac{\ln N}{\ln \langle k \rangle}$

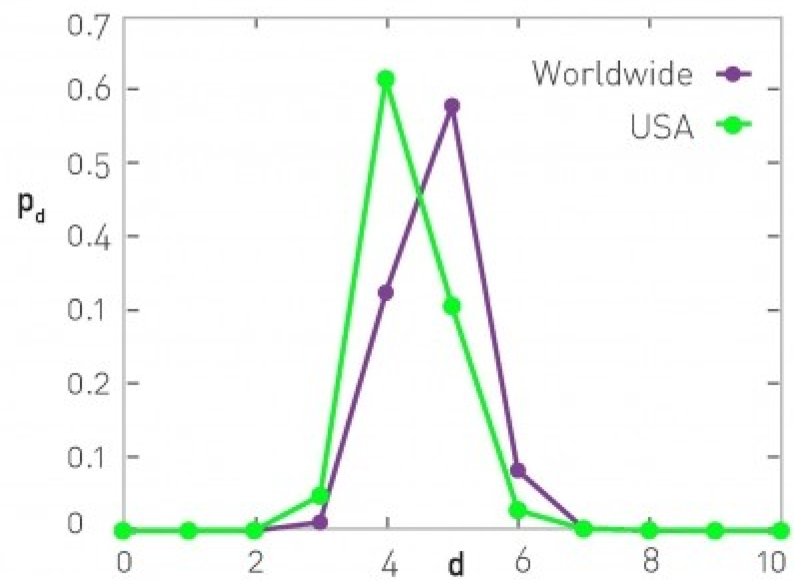
Network	N	L	‹k›	‹d›	d_{max}	ln*N*/ln‹*k*›
Internet	192,244	609,066	6.34	6.98	26	6.58
WWW	325,729	1,497,134	4.60	11.27	93	8.31
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile-Phone Calls	36,595	91,826	2.51	11.72	39	11.42
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	93,437	8.08	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

http://networksciencebook.com/chapter/3

Small Worlds

Left: Milgram's experiment. 64 of the 296 letters made it to the recipient.
Right: Distance distribution for all pairs of Facebook users.





http://networksciencebook.com/chapter/3

Clustering Coefficient

Are my friends also friends with each other?

$C_i = 0$ means that there are no links between *i*'s neighbors.
$C_i = 1$ implies that each of the *i*'s neighbors link to each other.

In a random network the probability that two of *i*'s neighbors link to each other is *p*. There are $k_i(k_i-1)/2$ possible links between them. The expected value of links is

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

Thus, the local clustering coefficient of a random network is independent of the node's degree:

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

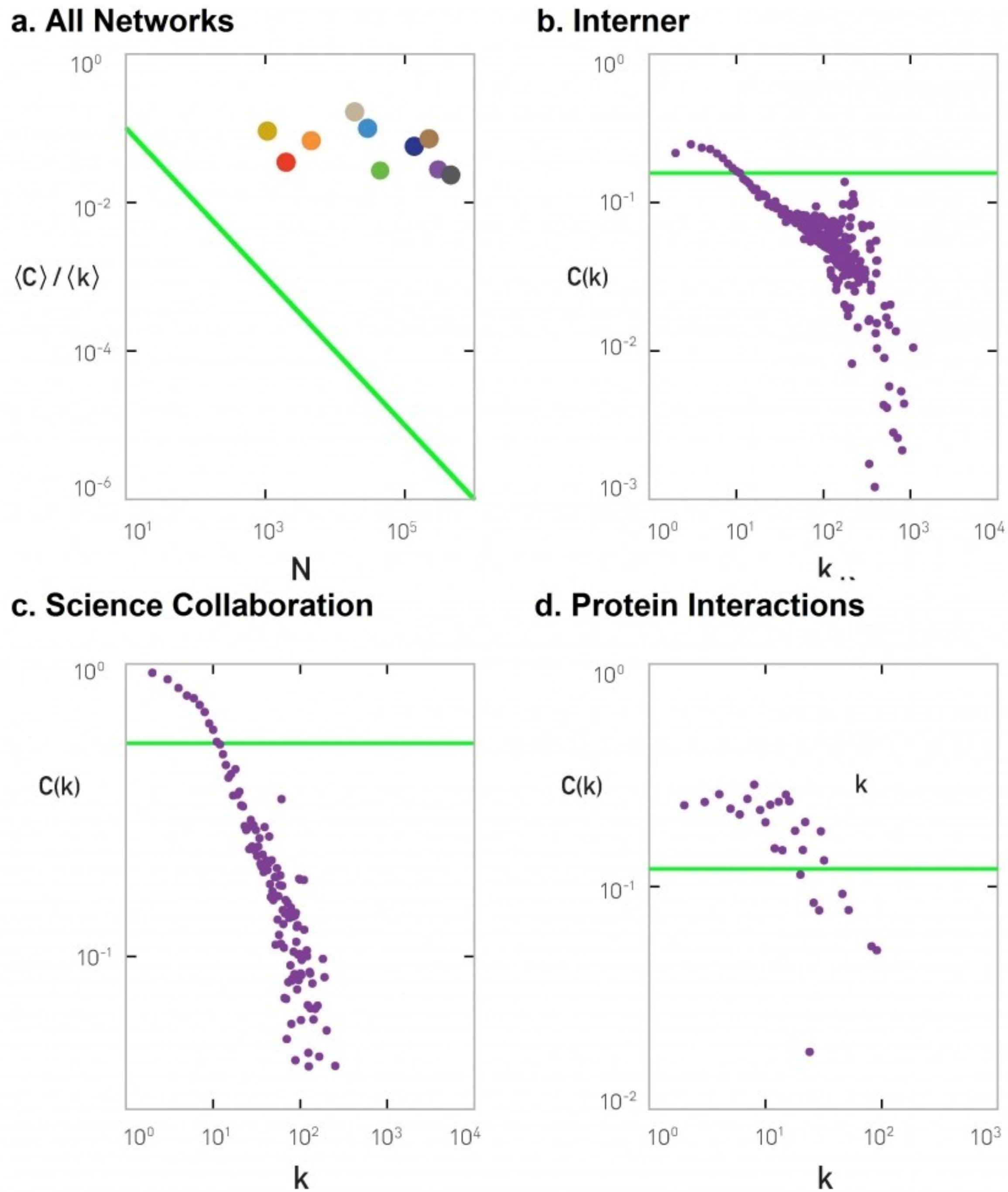
Clustering Coefficient

Real networks violate the Equation:

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

Real networks seems to have average clustering coefficient independent of N.

On the other hand, it depends on k.



http://networksciencebook.com/chapter/3

In summary

Random networks have several properties:

- Degree distribution: binomial, but well approximated by Poisson when $\langle k \rangle \ll N$.
- Connectedness: a giant component exists when $\langle k \rangle > 1$.
- Average Path Length: accounts for the emergence of small world phenomena. $$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$$
- Clustering Coefficient: it is independent of the node's degree and $\langle C \rangle$ depends on the system size as follows:

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

- Apparently, the small world phenomena is the only property reasonably explained by the ER model. Other characteristics of real world networks deviate from the ER model.

http://networksciencebook.com/chapter/3

In summary

Question: If real networks are not random, why are we studying the ER model?

Answer: If a property resembles the ER model, one could argue that it probably represents something that occurred by chance. Otherwise, it may present some signature of order, requiring a deeper explanation.

Useful code:
https://github.com/CambridgeUniversityPress/FirstCourseNetworkScience/blob/master/tutorials/Chapter%205%20Tutorial.ipynb

In summary

Box 3.11

At a Glance: Random Networks

Definition: N nodes, where each node pair is connected with probability p.

Average Degree:

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

Average Number of Links:

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

Average Degree:

Binomial Form:

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

Poisson Form:

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

Giant Component (GC) (N_G):

$\langle k \rangle < 1 : \quad N_G \sim \ln N$

$1 < \langle k \rangle < \ln N : \quad N_G \sim N^{\frac{2}{3}}$

$\langle k \rangle > \ln N : \quad N_G \sim (p - p_c) N$

Average Distance:

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

Clustering Coefficient:

$\langle C \rangle = \frac{\langle k \rangle}{N}$

http://networksciencebook.com/chapter/3