Random Network Properties

Introduction to Network Science

Instructor: Michele Starnini — https://github.com/chatox/networks-science-course

Universitat Pompeu Fabra Barcelona

Contents

- Connectedness under the ER model
- Distances under the ER model
- Clustering coefficient under the ER model

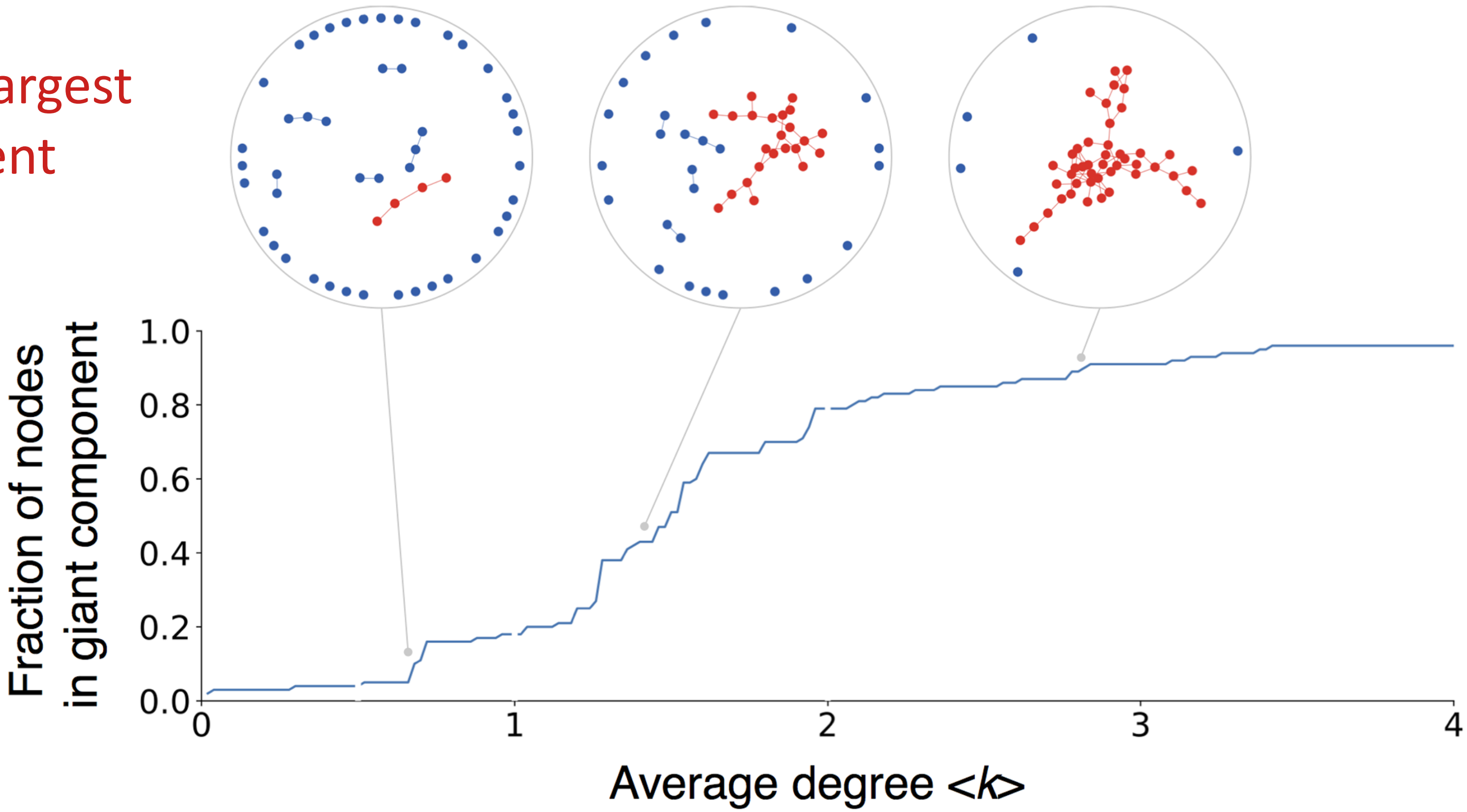
Connectivity in ER networks

ER network model

- Has two parameters: G(p,N)
- For p << 1 and N >> 1, there is a single parameter, <k> (completely determined by p)
- How does the connectivity depend on <k>?
- How big is the largest connected components?

An interesting property of ER networks

Red = nodes in the largest connected component



Source: Menczer, Fortunato, Davis: A First Course on Networks Science. Cambridge, 2020.

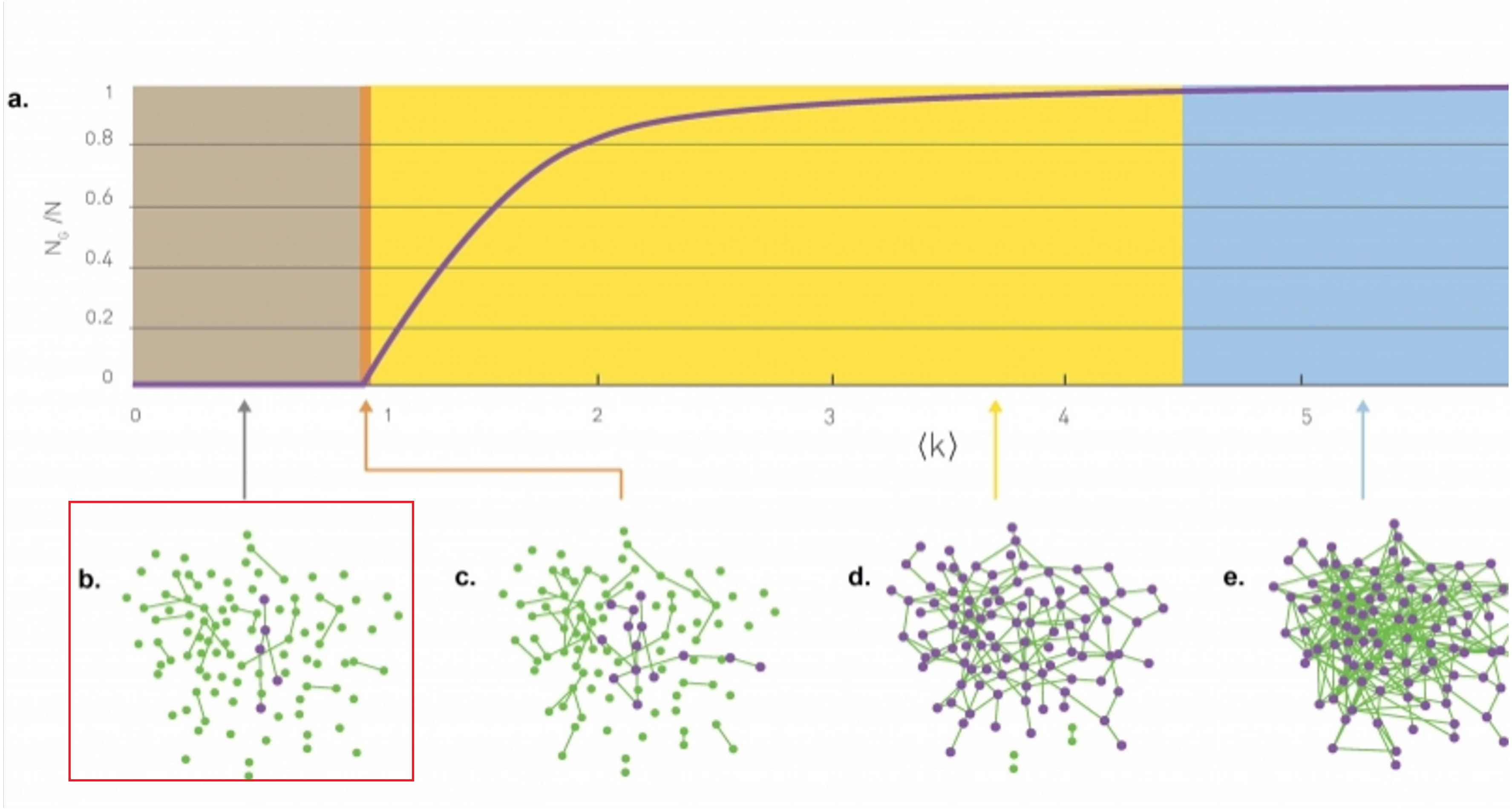
ER network as <k> increases

- When <k> = 0: only singletons
- When <k> < 1: several small disconnected components
- When <k> > 1: giant component: “many” nodes are connected
- When <k> = N – 1 complete graph

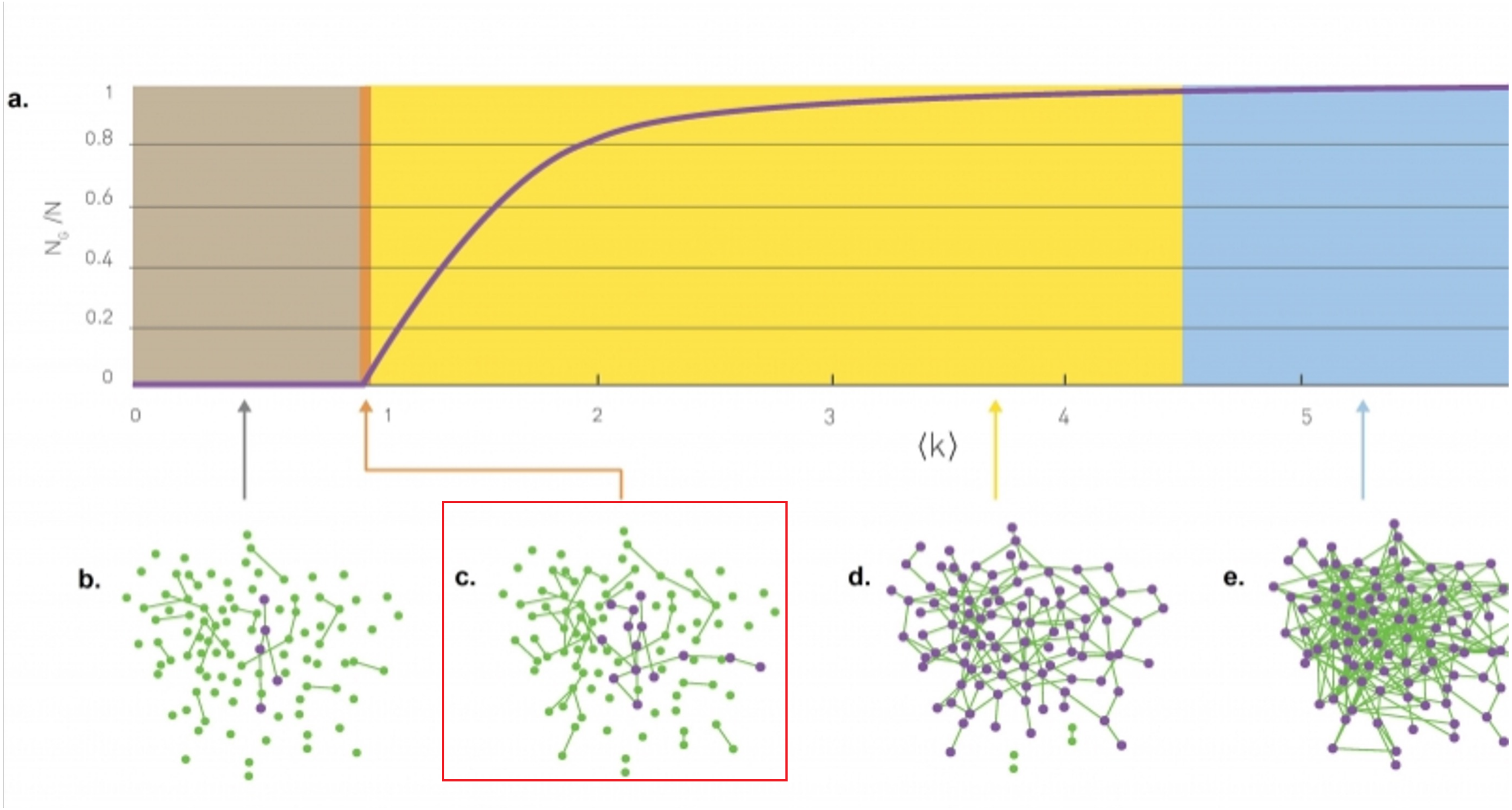
to have a giant comp it is **necessary** that <k> = 1 (at least one link per node)
Erdös and Rényi proved it is **sufficient** in 1959

This result holds **on average**, not on every execution of the model

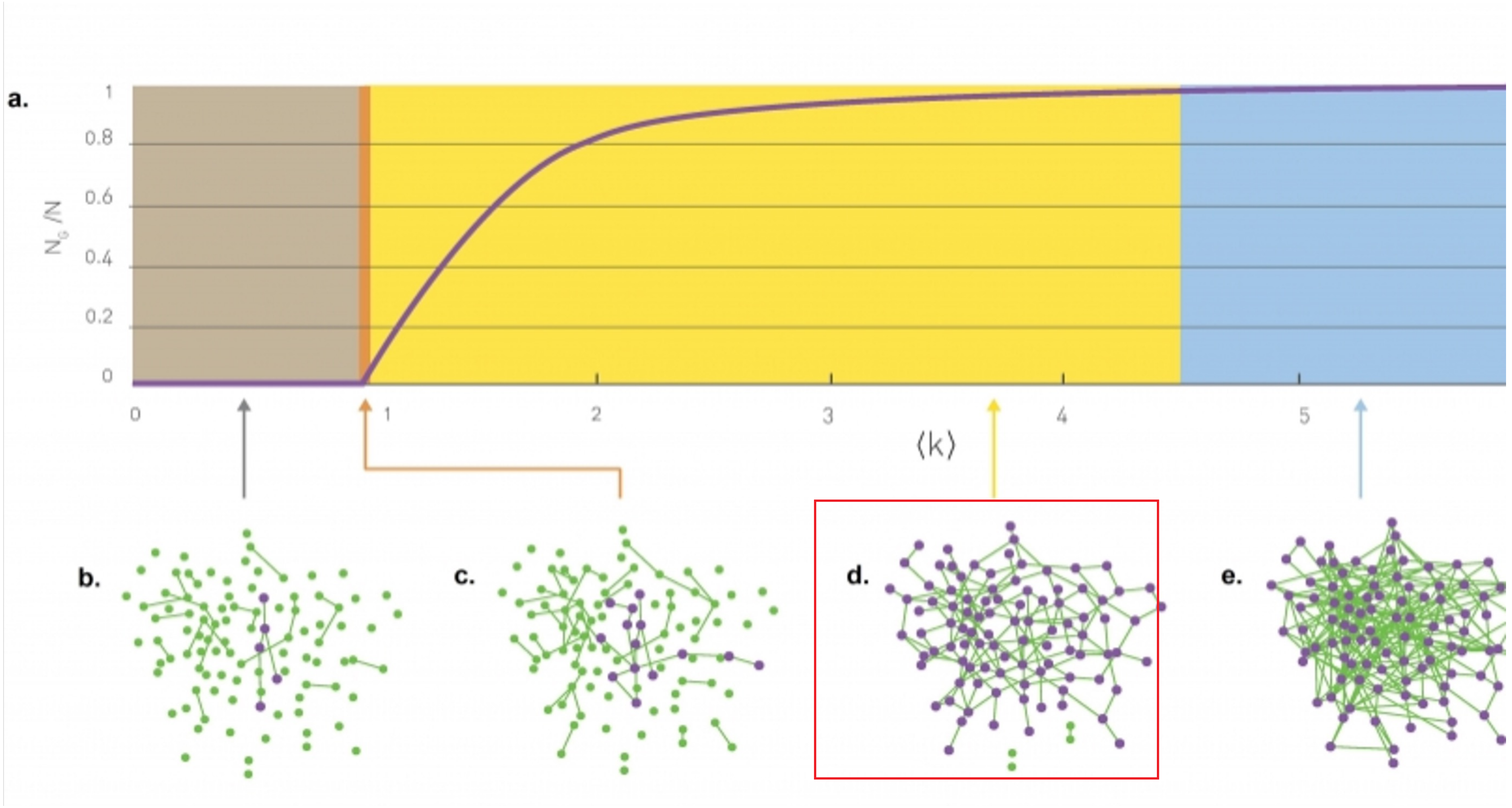
Sub-critical regime $\langle k \rangle < 1$



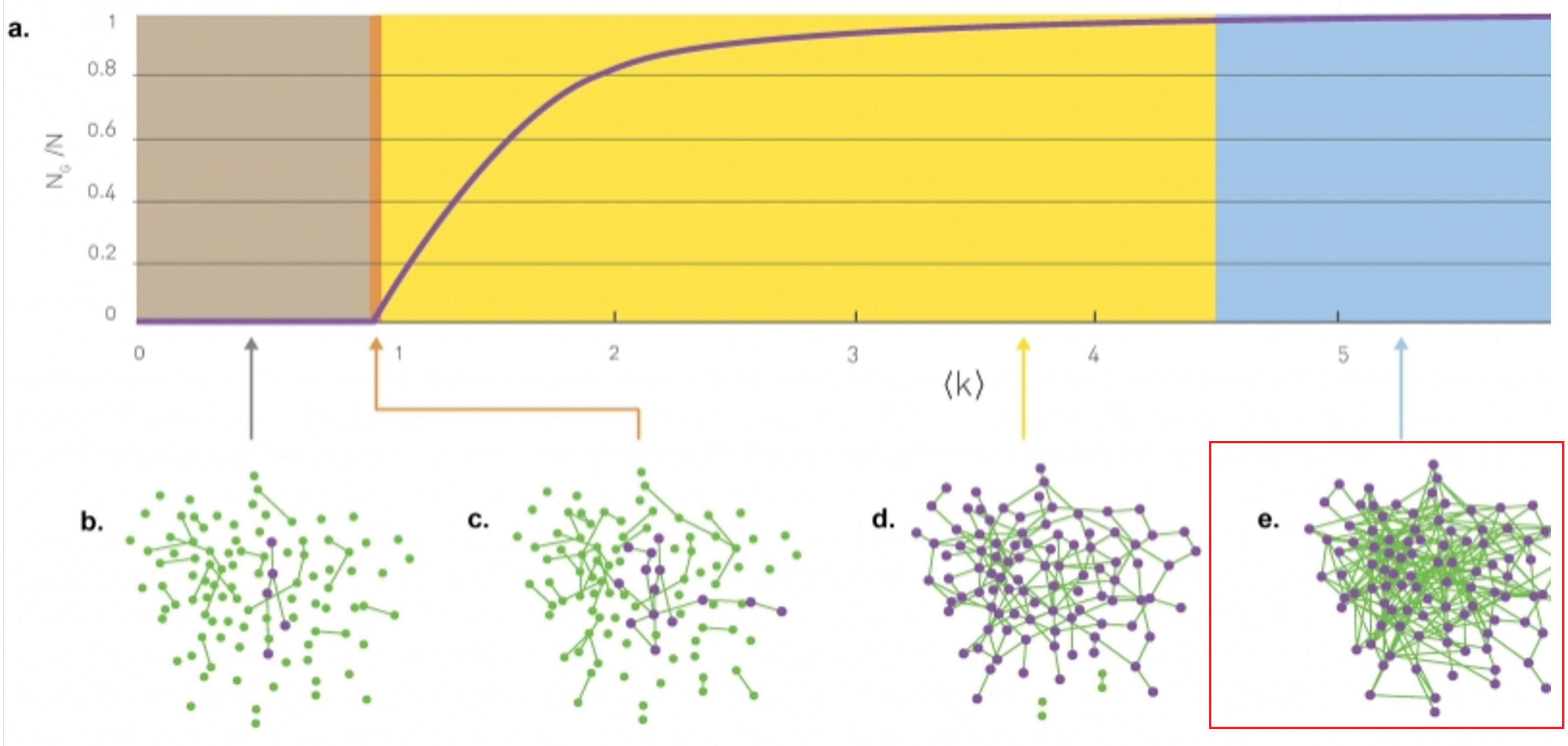
Critical point: $\langle k \rangle = 1$



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



Connected regime: $\langle k \rangle > \log N$



Disconnected nodes to giant component

- The size of the largest component does not increase “smoothly” with $<k>$
- There is an **abrupt** increase (on average) for $<k>=1$
- $<k>_c=1$ is a **critical point** (value) separating two regimes
- The larger the network, the more abrupt the change!

$<k>_c=p_c(N-1)$, $\mathbf{p_c=1/N}$

Very large network are (almost) always connected, no matter how small is p!

Phase transitions

- ER is a "static" network model: from <k> to a graph
- Consider dynamics: you slightly increase <k> from 0 to N-1, at a certain value of <k>, a giant component **emerges**
- Equivalent to increasing temperature of water: at a certain temperature (100 C), water becomes steam
- (liquid) water & steam are two different **phases** of the same thing (water). The "critical" temperature T_c=100 determines the "**phase transition**".
- The "critical" connectivity $\langle k \rangle_c$=1 determines the phase transition between a disconnected and (almost) connected network (existence of a giant component)
- Why "almost"? how many connected nodes are enough for a "giant" component?

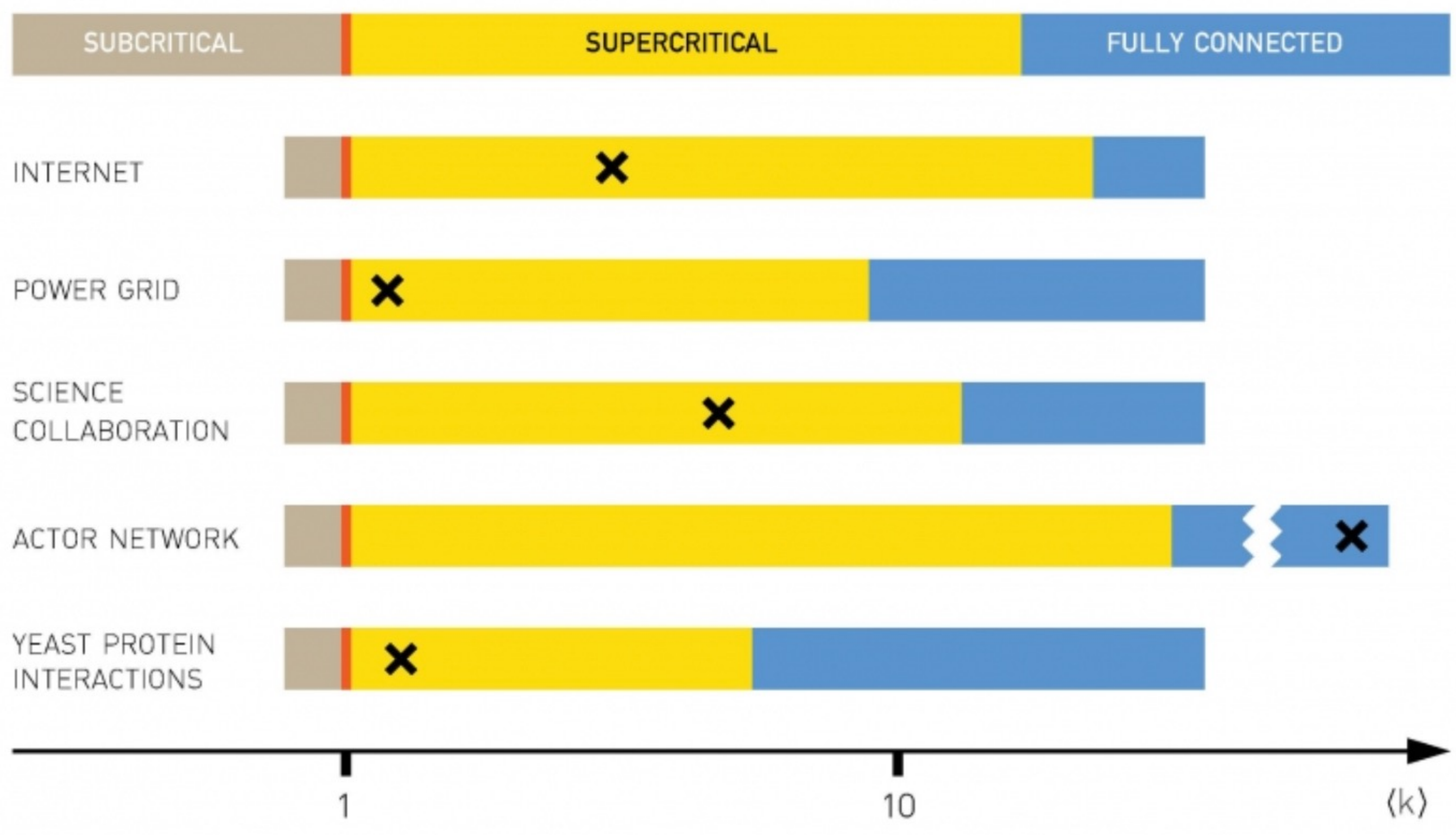
Most real networks are supercritical:

$$\langle k \rangle > 1$$

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

Most real networks are supercritical:

$$\langle k \rangle > 1$$



Small-world phenomenon
a.k.a. “six degrees of separation”

"Small-world phenomenon"

- If you choose any two individuals on Earth, they are connected by a relatively short path of acquaintances
- Formally
  - The expected distance between two randomly chosen nodes in a network grows much slower than its number of nodes

How many nodes at distance ≤d?

In an ER graph:

nodes at distance 1 $\langle k \rangle$

nodes at distance 2 $\langle k \rangle^2$

…

nodes at distance d $\langle k \rangle^d$

Max number of nodes… is N ☺

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

What is the maximum distance?

- Assuming $\langle k \rangle \gg 1$ $\quad \mathrm{N(d_{max})} = \frac{\langle k \rangle^{d_{\max}+1} - 1}{\langle k \rangle - 1} \approx N$

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

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

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

Empirical average and maximum distances

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

Why?

- Heterogeneity: the probability distribution of distances d is heterogeneous. There are only a few long paths (so large d_{max}), many short ones.

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

Clustering coefficient

or

”a friend of a friend is my friend”

Clustering coefficient C_i of node i

- Remember
  - $C_i = 0 \Rightarrow$ neighbors of *i* are disconnected
  - $C_i = 1 \Rightarrow$ neighbors of *i* are fully connected

Links between neighbors in ER graphs

- The number of nodes that are neighbors of node i is k_i
- The number of distinct pairs of nodes that are neighbors of i is $k_i(k_i-1)/2$
- The probability that any of those pairs is connected is p
- Then, the expected links L_i between neighbors of *i* are:

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

Clustering coefficient in ER graphs

- Expected links L_i between neighbors of i:

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

- Clustering coefficient

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

Very small for large graphs!

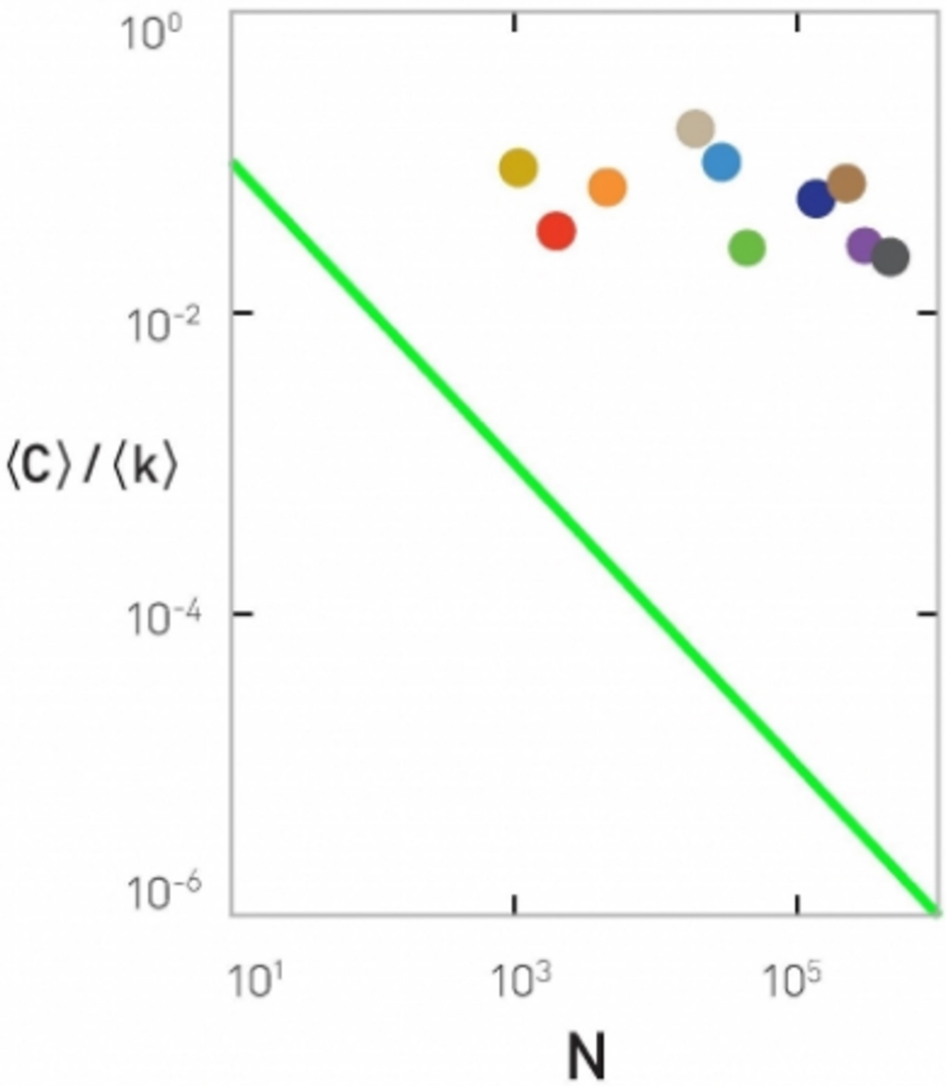
$$= p \approx \frac{\langle k \rangle}{N}$$

In an ER graph

$$C_i = \langle k \rangle / N$$

If ⟨k⟩ is fixed, large networks should have smaller clustering coefficient

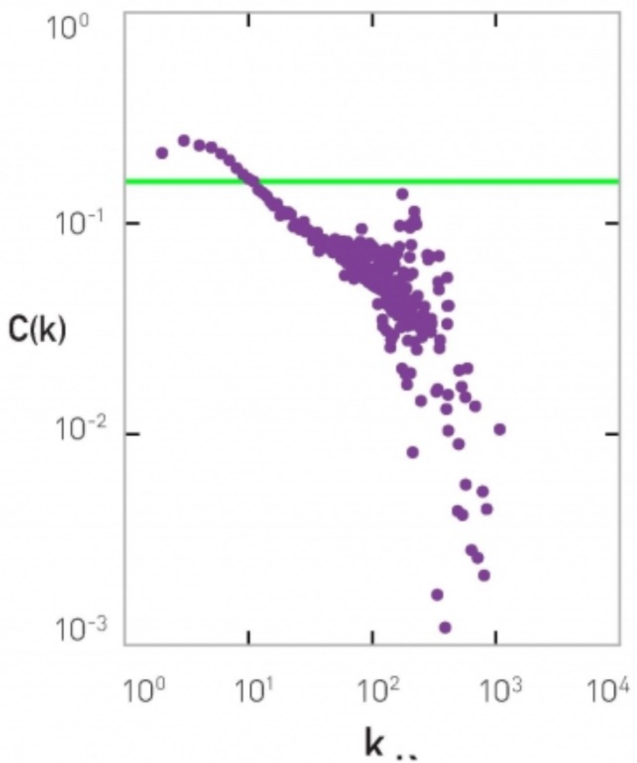
We should have that *⟨C⟩/⟨k⟩* follows *1/N*



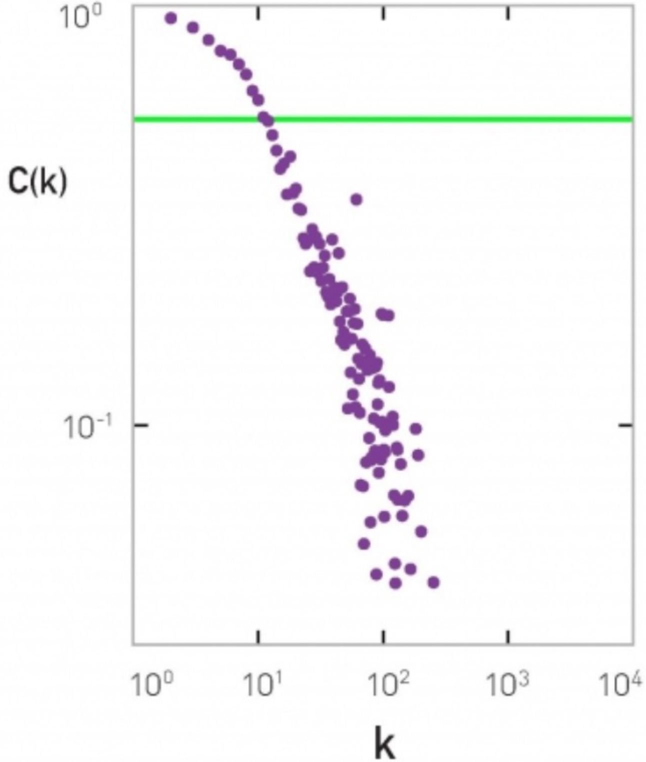
In an ER graph $C_i = \langle k \rangle / N$

Clustering should be independent of the degree

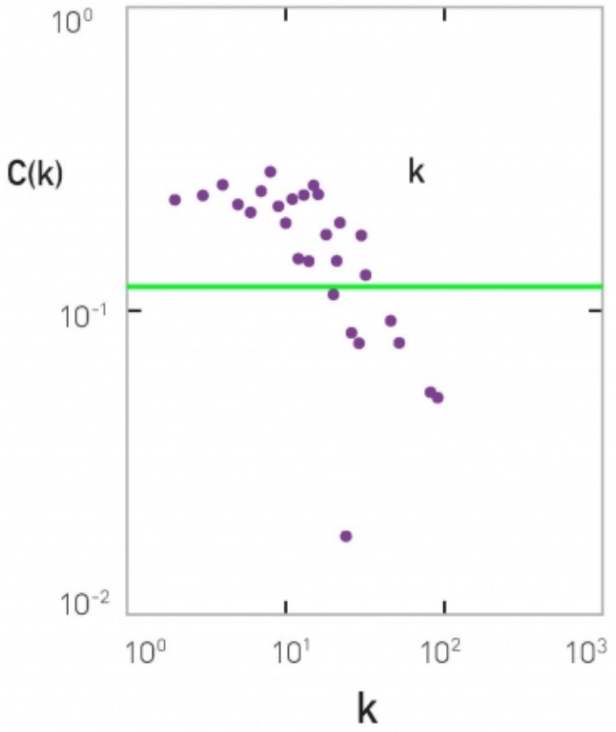
Internet



Science collaborations



Protein interactions



To re-cap …

ER network is a bad model for **degree distribution**

- Predicted

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

- Observed

Many nodes with larger degree than predicted

ER network is a good model of **path length**

- Predicted

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

- Observed

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

‹d›	d_{max}	ln*N*/ln‹*k*›
6.98	26	6.58
11.27	93	8.31
18.99	46	8.66
11.72	39	11.42
5.88	18	18.4
5.35	15	4.81
3.91	14	3.04
11.21	42	5.55
2.98	8	4.04
5.61	14	7.14

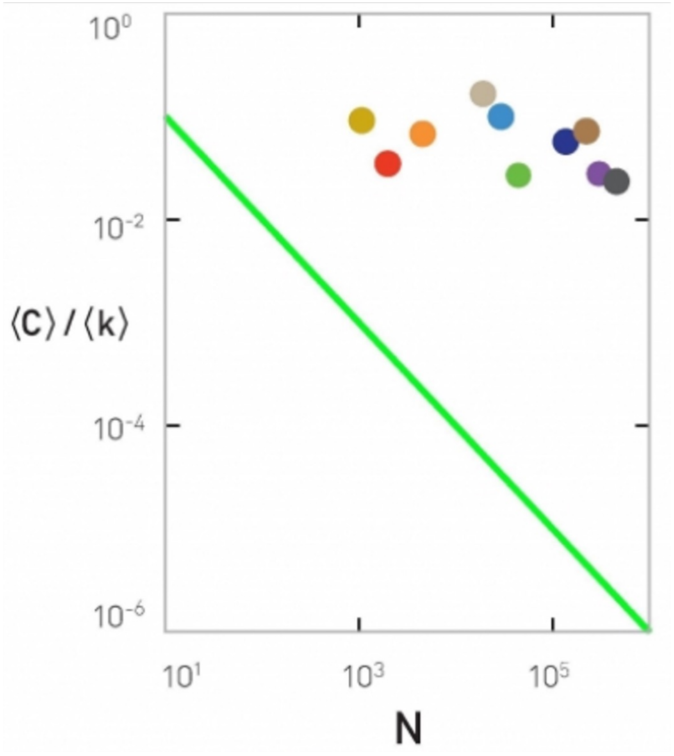
ER network is a bad model of **clustering coefficient**

- Predicted

$$C_i = \langle k \rangle / N$$

- Observed

Clustering coefficient decreases if degree increases



Why do we study the ER model?

- Starting point
- Simple
- Instructional
- Historically important, and gained prominence only when large datasets started to become available $\Rightarrow$ relevant to Data Science!

Exercise [B. 2016, Ex. 3.11.1]

Consider an ER graph with $N=3{,}000$ $p=10^{-3}$

1)<k> $\simeq$?

2)In which regime is the network?

$$\langle k \rangle < 1, \langle k \rangle = 1, \langle k \rangle > 1, \langle k \rangle > \log N$$

3)Suppose we want to increase N until there is <u>only one connected component</u>
3.1) What is <k> as a function of p and N?
3.2) What should N be, then? Let's call that value N^{cr}
Write the equation and solve by trial and error

$$\langle k \rangle \approx \log N$$

4)What is <k> if the network has N^{cr} nodes?

5)What is the expected distance <d> with N^{cr} nodes?

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

Summary

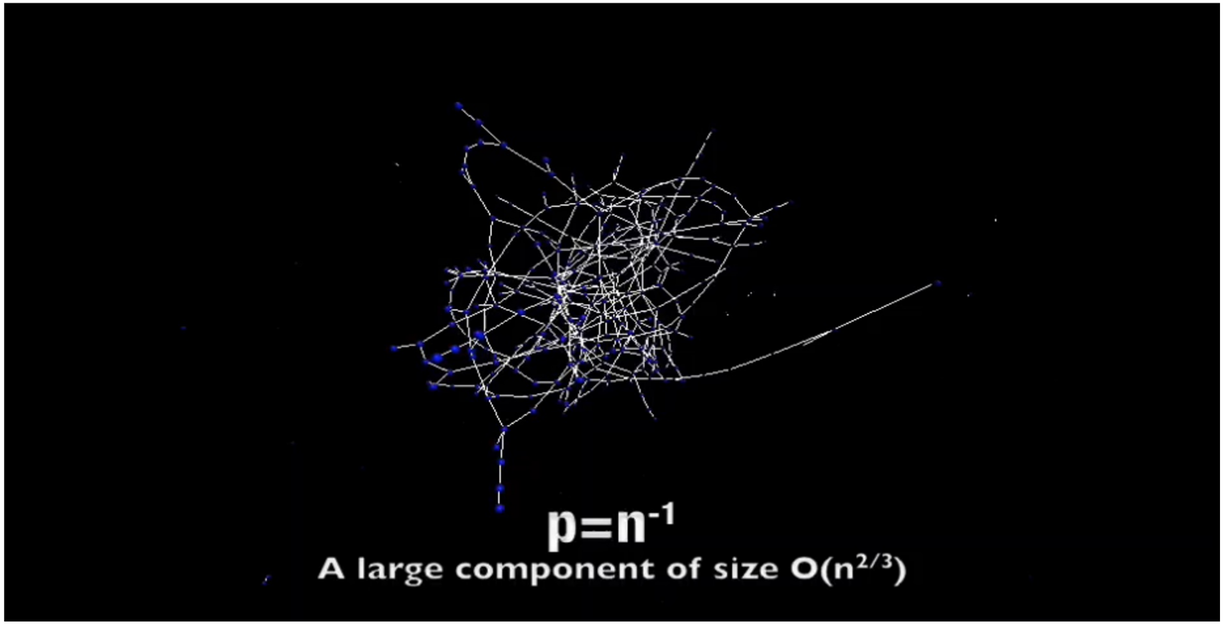
Things to remember

- The ER model
- Degree distribution in the ER model
- Distance distribution in the ER model
- Connectivity regimes in the ER model

Practice on your own

- Take an existing network
  - (e.g., from the slide "Empirical average and maximum distances")
  - Assume it is an ER network
  - Indicate in which regime is the network
  - Estimate expected distance
  - Compare to actual distances, if available
- Write code to create ER networks

Another visualization of the emergence of a giant connected component



http://networksciencebook.com/images/ch-03/video-3-2.m4v

Sources

- A. L. Barabási (2016). Network Science – Chapter 03
- Data-Driven Social Analytics course by Vicenç Gómez and Andreas Kaltenbrunner
- URLs cited in the footer of specific slides