Properties of Random Networks

Social Networks Analysis and Graph Algorithms

Prof. Carlos Castillo — https://chato.cl/teach

Universitat Pompeu Fabra Barcelona

Contents

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

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

The "Magtension" game

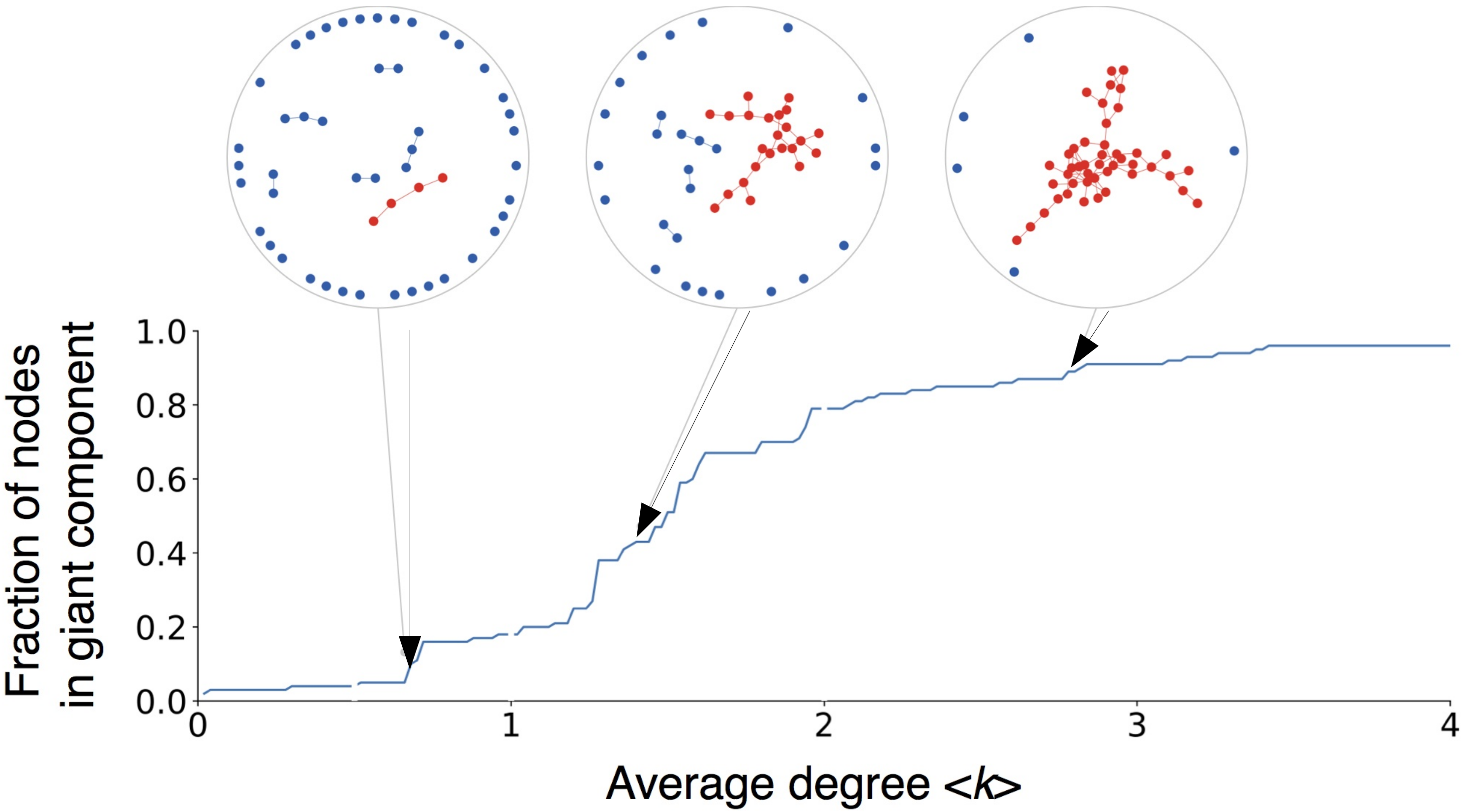
- Take turns placing one magnet inside an enclosed space
- You lose if, after your play, any two magnets stick to each other

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

Connectivity in ER networks

An interesting property of ER networks

Red = nodes in largest connected component



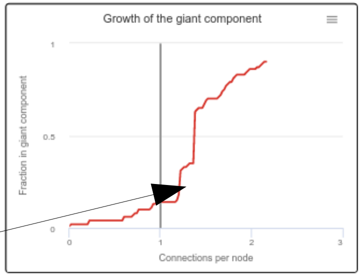
Source: Menczer, Fortunato, Davis: A First Course on Networks Science. Cambridge, 2020.

Exercise

Giant component under ER

Go to netlogoweb.org/launch and select:
"Sample Models / Networks / Giant component"

- Execute the "Giant Component" program in Netlogo Web
  - Select num-nodes N (e.g., 100)
  - Click "setup"
  - Click "go"
  - Write down the point at which there is an *elbow* in the distribution of links
  - Repeat various times



- **Indicate approximately where, on average, you find the "elbow"**

Pin board: https://upfbarcelona.padlet.org/chato/jk9oinetdyuzecol

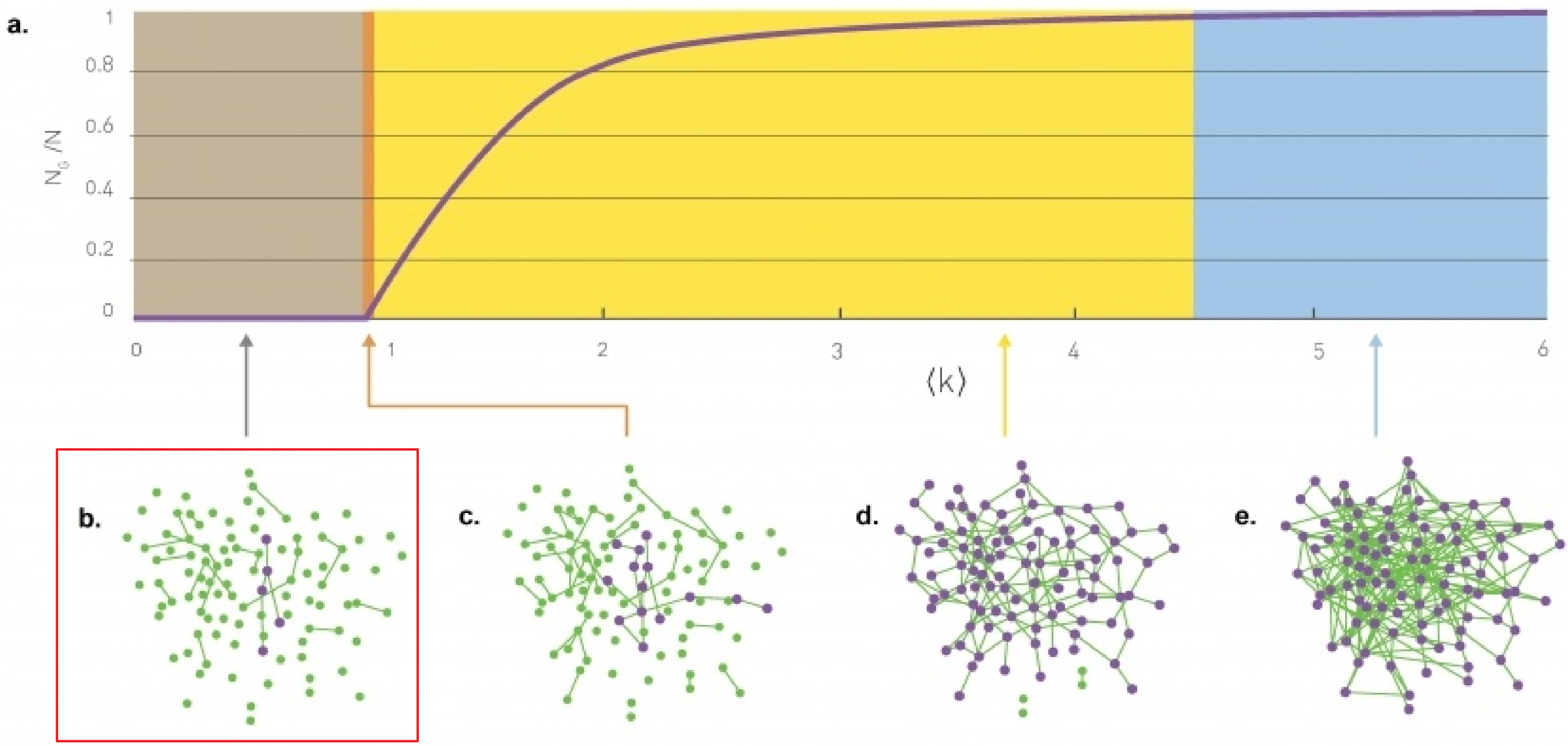
ER network as $<k>$ increases

- When $<k> = 0$: only singletons
- When $<k> < 1$: disconnected
- When $<k> > 1$: giant connected component
- When $<k> = N - 1$ complete graph

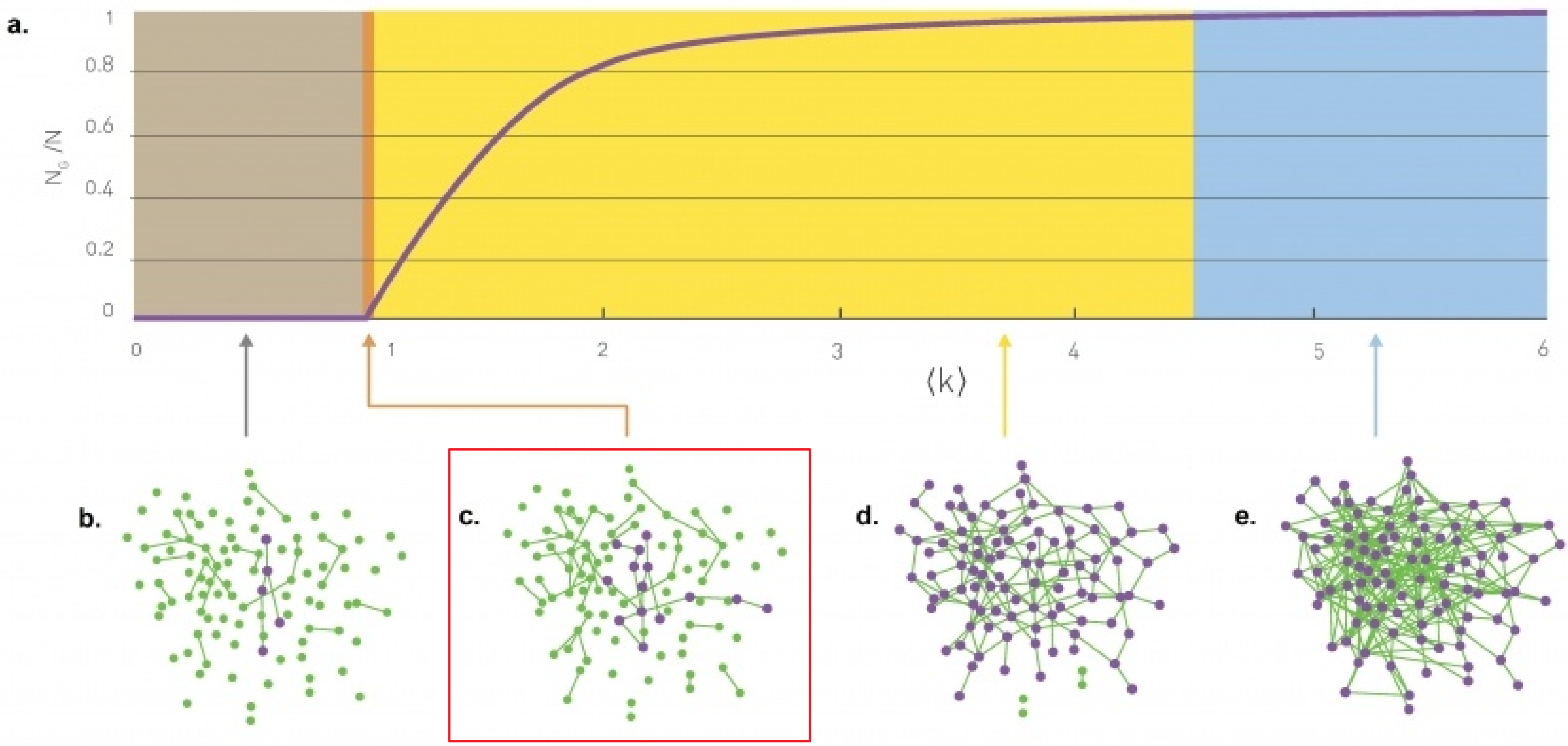
It's obvious that to have a giant connected it is **necessary** that $<k> = 1$
Erdös and Rényi proved it is **sufficient** in 1959

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

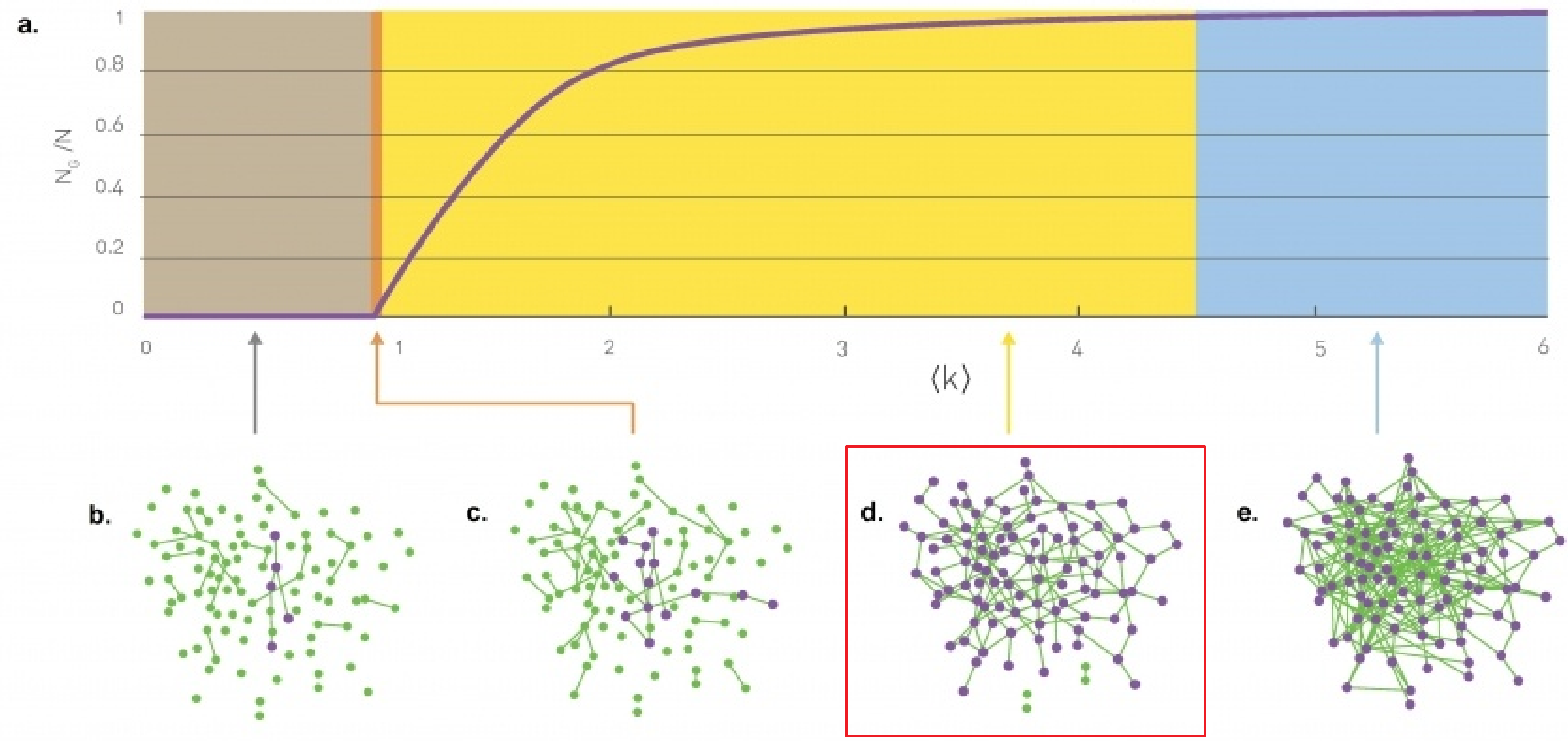
Sub-critical regime: $\langle k \rangle < 1$



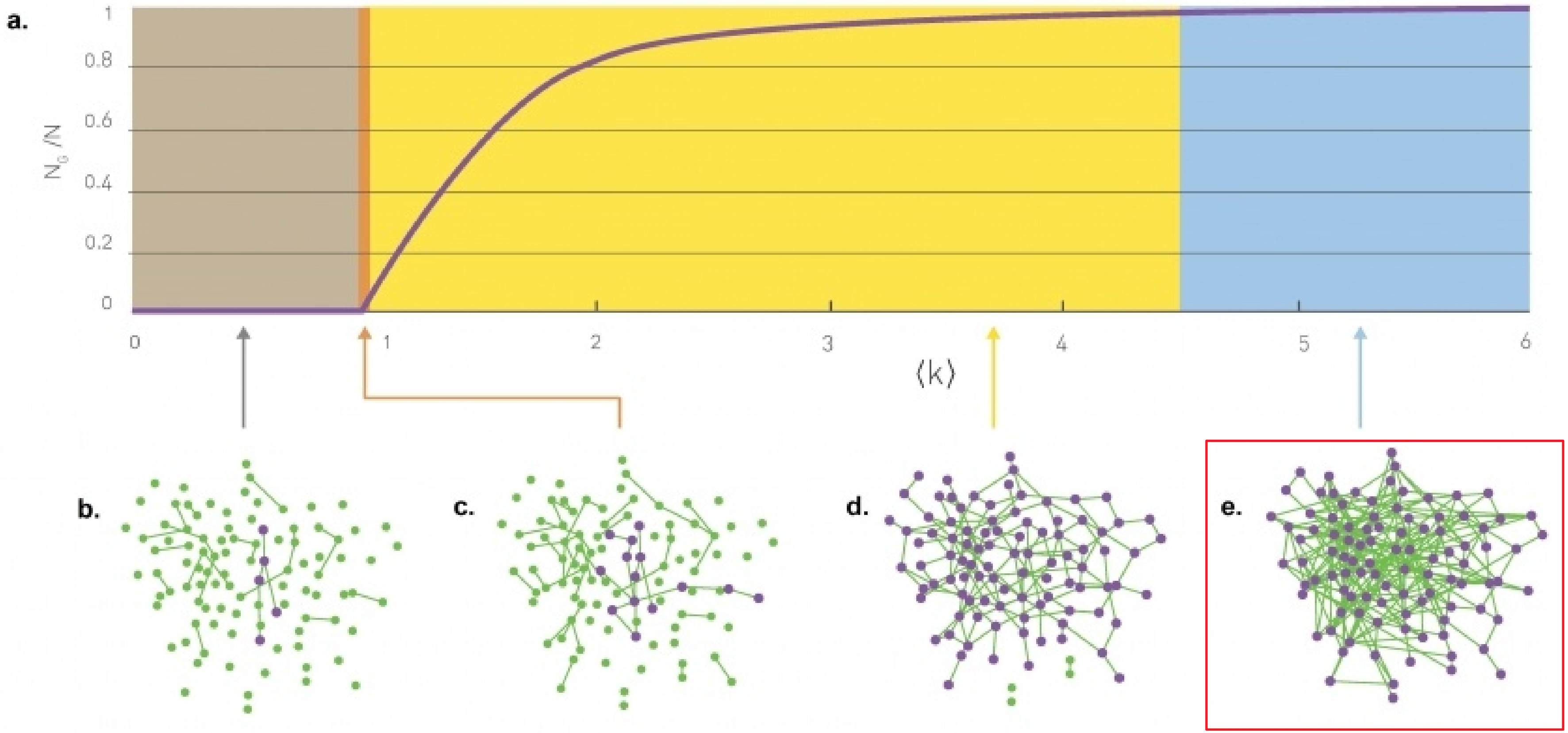
Critical point: $\langle k \rangle = 1$



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



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



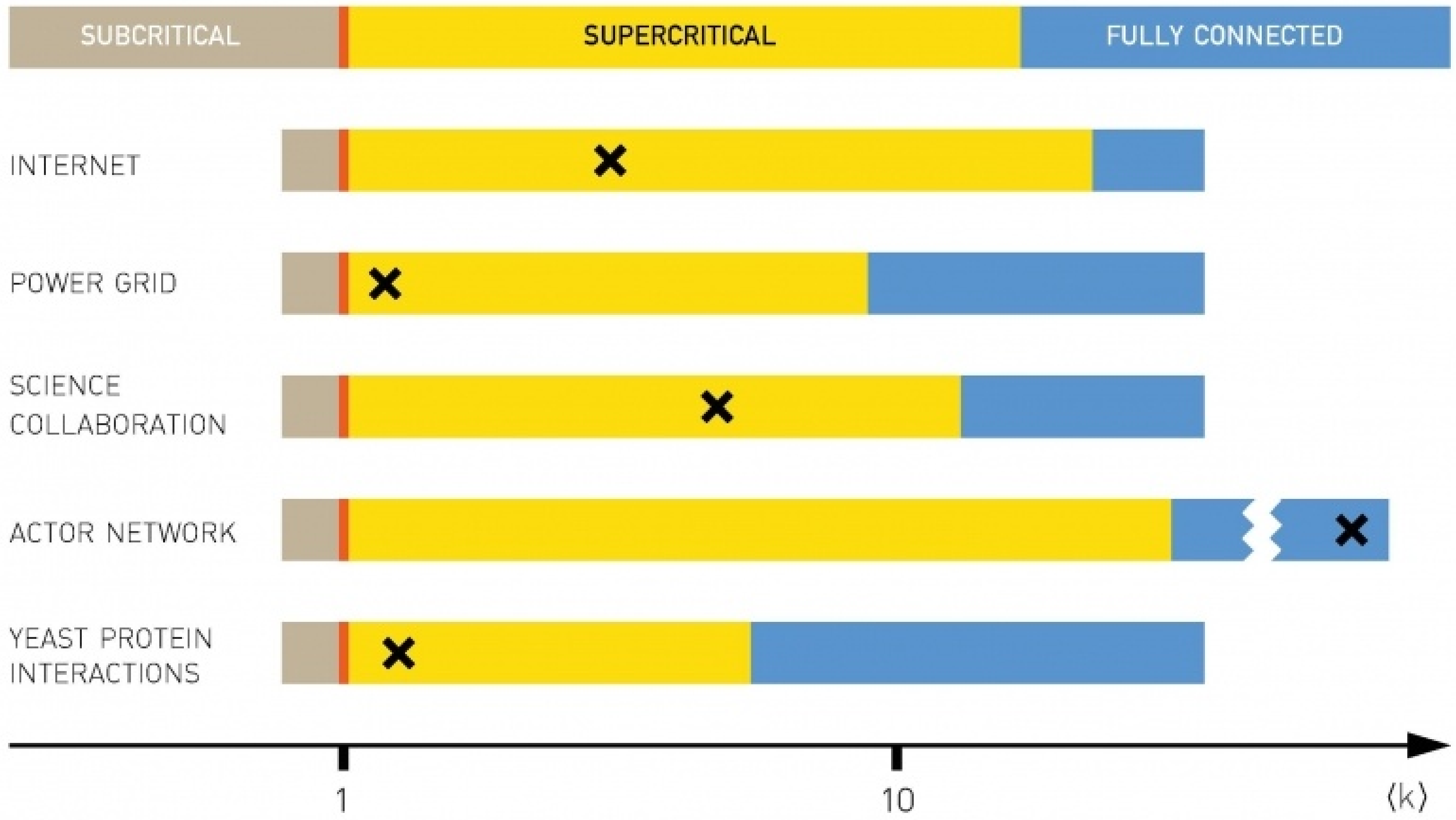
Most real networks are supercritical:

$$\langle k \rangle > 1$$

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:

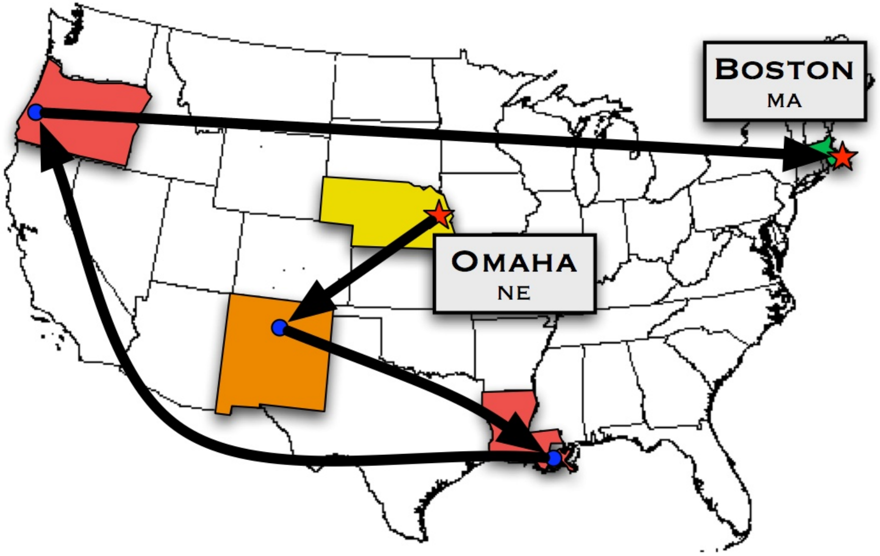
$$\langle k \rangle > 1$$



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

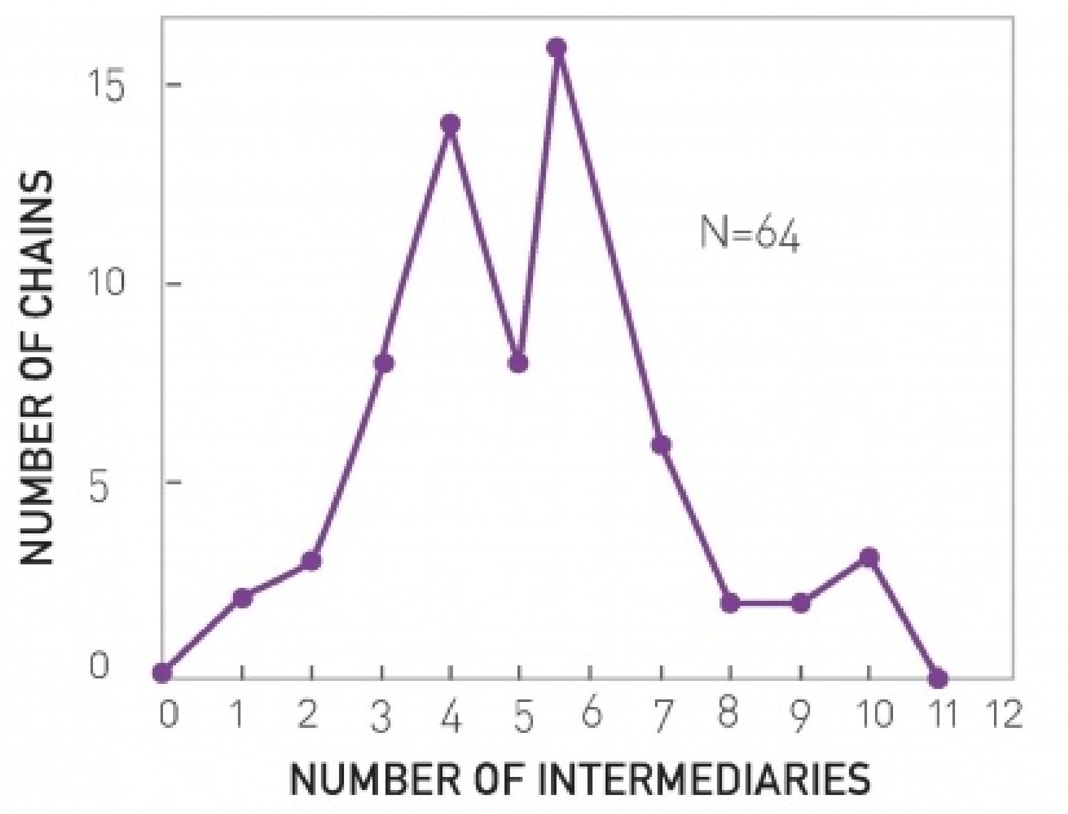
Milgram's experiment in 1967

- Instructions: send to personal acquaintance most likely to know the target
  - Sources: 160 people in Wichita and Omaha
  - Targets: (1) a stock broker in Boston, MA and (2) a student in Sharon, MA
- Materials: short summary of study purpose, target photograph, name, address and information



Milgram's experiment in 1967 (results)

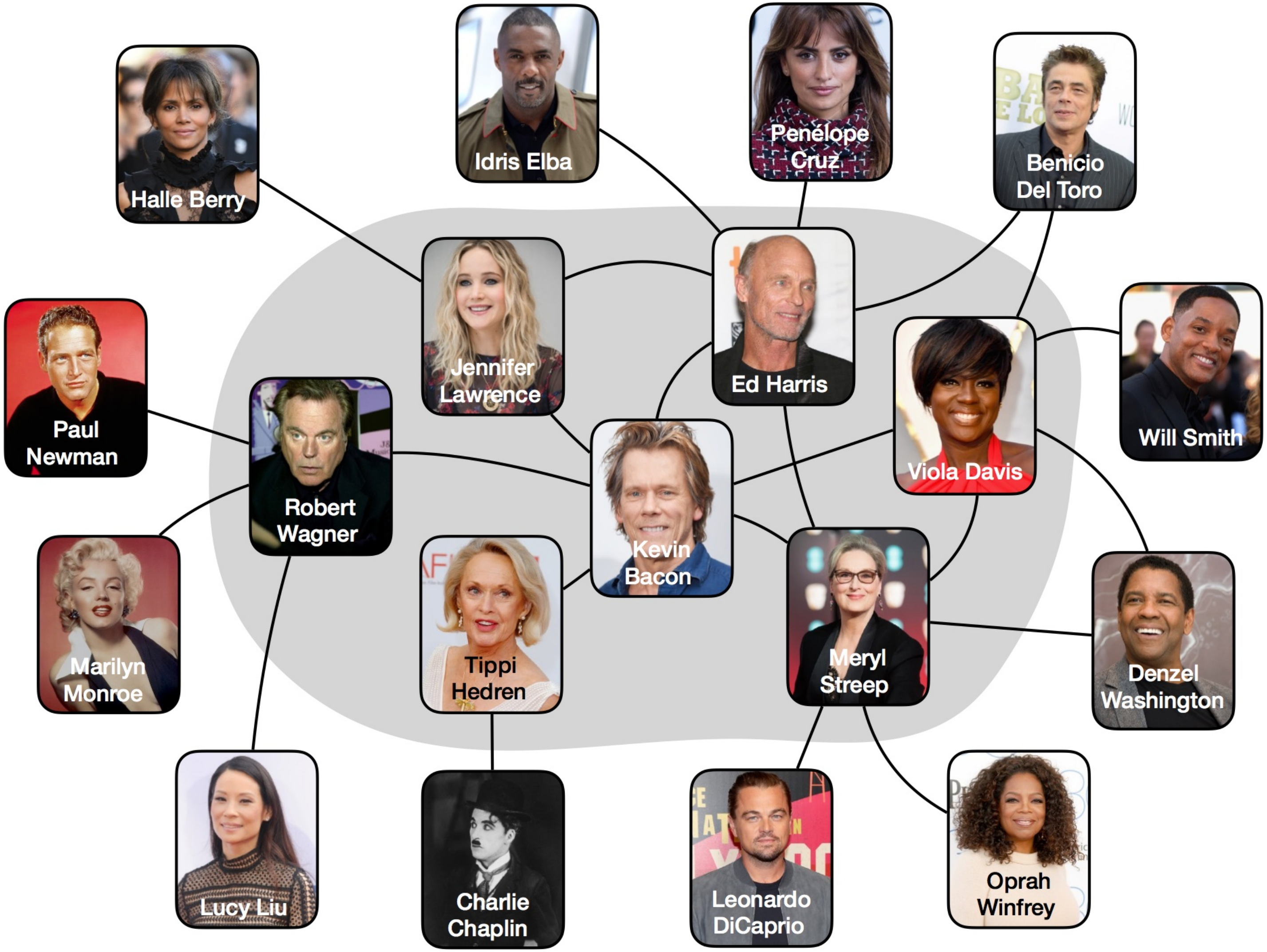
- **<u>64 of 296</u> (22%) of the letters reached their destination**
- Average 6.5 steps, much lower than expected



Wikipedia Speedruns

- Select Wikipedia's "Random article" twice
- Go from one to the other only by clicking links; no "Ctrl-F" search allowed
- Timeout at 30 seconds
- Example: from *John Cena* to *Double-stranded RNA viruses*

https://www.youtube.com/shorts/5qdZxREcCWw



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

https://oracleofbacon.org/

Stormy Daniels has a Bacon number of 2. [Find a different link]

Stormy Daniels

was in

The 40 Year Old Virgin (2005)

with

Charlie Hartsock

was in

Crazy, Stupid, Love. (2011)

with

Kevin Bacon

[Kevin Bacon] to [Stormy Daniels] [Find link] [More options >>]

“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:

$\langle k \rangle$ nodes at distance 1

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

...

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

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

What is the maximum distance?

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

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

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

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

Empirical average and maximum distances

Network	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\ln N / \ln \langle k \rangle$
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	93,437	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

Approximation

- Given that d_{max} is dominated by a few long paths, while $<d>$ is averaged over all paths, in general we observe that in an ER graph:

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

Simple Exercise

Find a famous actress/actor far from Kevin Bacon

Go to https://oracleofbacon.org/ and find a famous actress or actor that has a distance from Kevin Bacon larger than

$$\langle d \rangle \approx \frac{\log N}{\log \langle k \rangle} = \frac{\log 702388}{\log 83.71} \approx 3$$

Write the name of the actress/actor and its distance

Tip: first look for some list of famous actresses/actors

Pin board: https://upfbarcelona.padlet.org/chato/rh0y9s1vngxk7klv

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 $k_i(k_i\text{-}1)/2$
- 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: $\langle L_i \rangle = p\dfrac{k_i(k_i-1)}{2}$

- Clustering coefficient

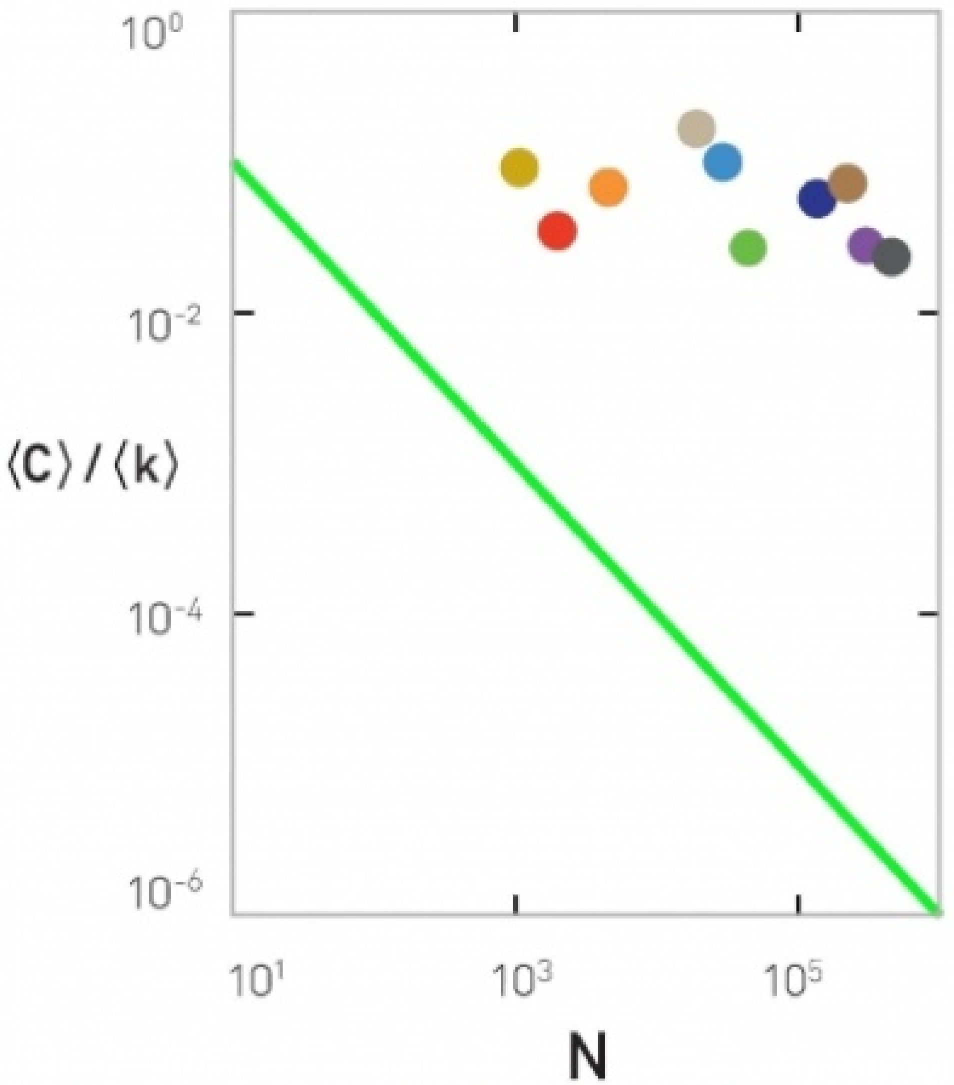
$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i-1)}$$
$$= \frac{2p\frac{k_i(k_i-1)}{2}}{k_i(k_i-1)} = \frac{\langle k \rangle}{N}$$

In an ER graph

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

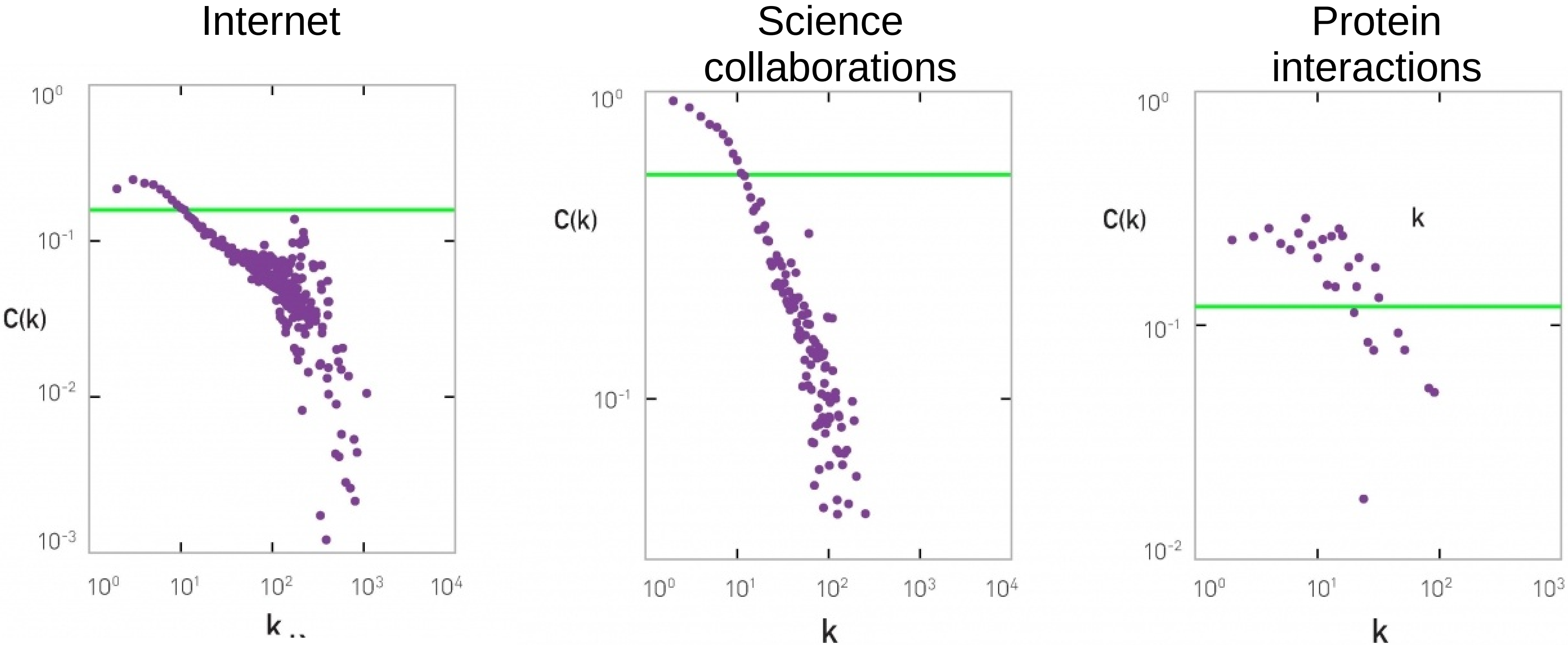
If $<$k$>$ is fixed, large networks should have smaller clustering coefficient

We should have that $<$C$>$/$<$k$>$ follows 1/N



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

Then the clustering coefficient of a node should be independent of the degree



To re-cap ...

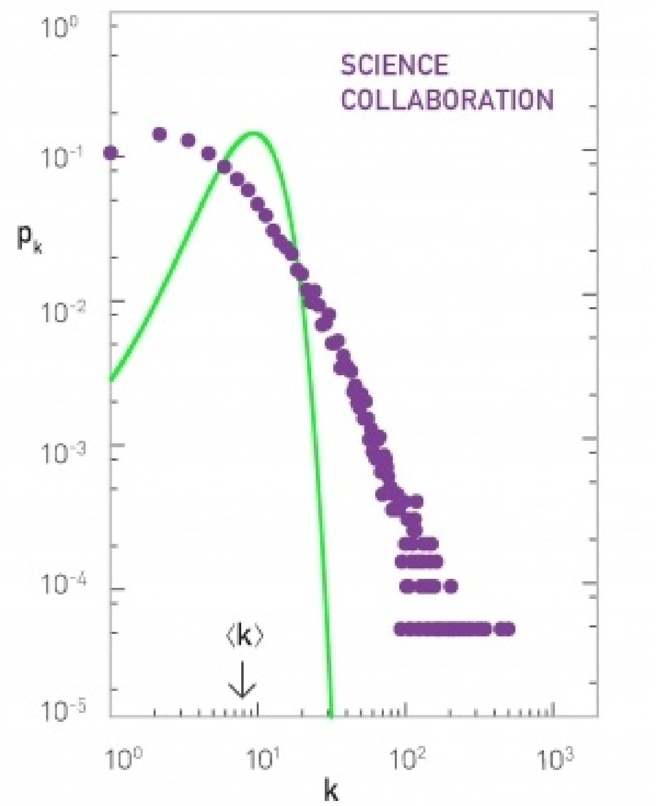
The ER model is a bad model of degree distribution

- Predicted

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

- Observed

  Many nodes with larger degree than predicted



The ER model is a good model of path length

- Predicted

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

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

‹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

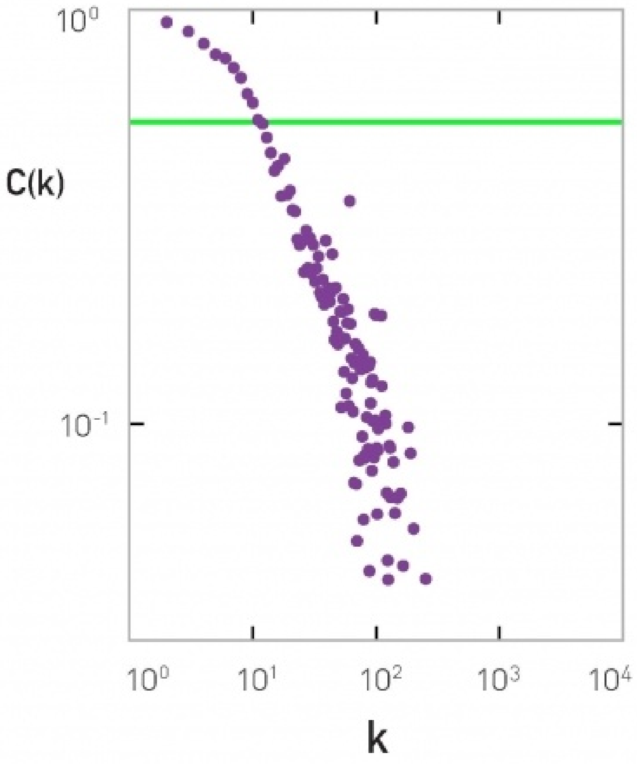
The ER model is a **bad model** of clustering coefficient

- 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> ≃ ?

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

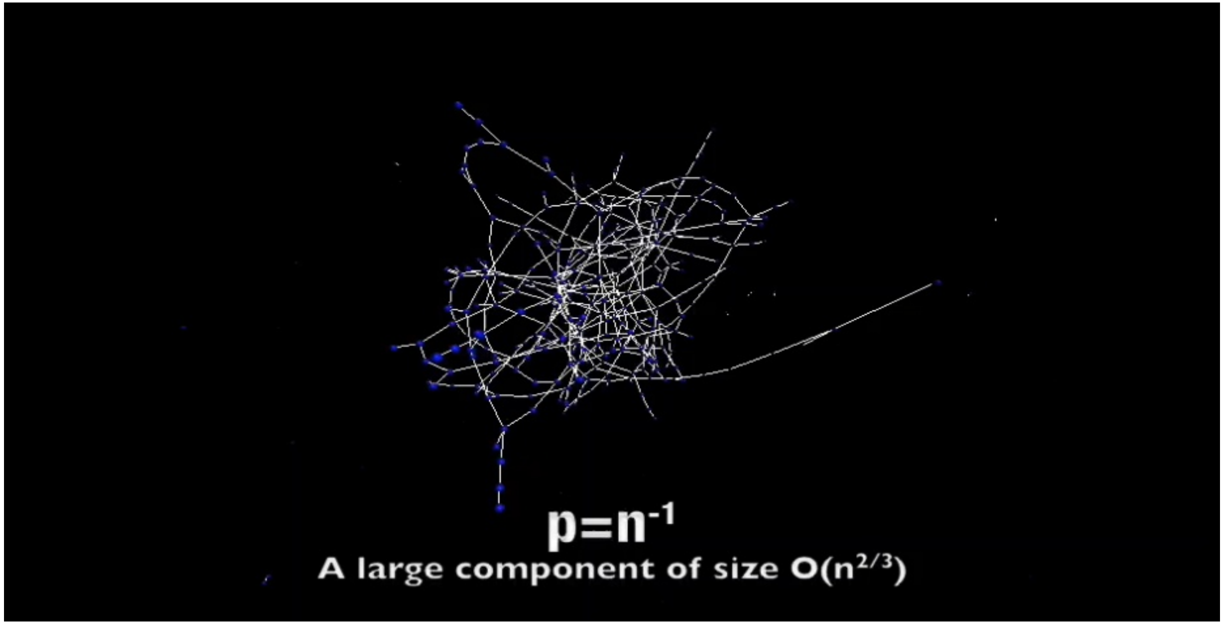
- The ER model
- Degree distribution in the ER model
- Distance distribution in the ER model
- Connectivity regimes in the ER model

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

Additional contents

Another visualization of the emergence of a giant connected component



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