Random networks

Introduction to Network Science

Carlos Castillo

Topic 03

Universitat Pompeu Fabra Barcelona

Sources

- Albert László Barabási: Network Science. Cambridge University Press, 2016.
  - Follows almost section-by-section chapter 03
- URLs cited in the footer of specific slides

Why studying random networks?

- One way of studying complex networks is by running stochastic models of network creation and then see if they generate networks that look like real ones
- The “random network” model is one such stochastic model in which each link is created independently at random

Meeting people at a party

- You pick a random person
- Talk to that person for a while, now you are connected
- Then pick another person
  - And repeat
- The result is what we call a **random network**

Formalization (Erdös-Rényi or ER)

- For each pair of nodes in the graph
  - Perform a Bernoulli trial with probability p
    - “Toss a biased coin with probability p of landing heads”
  - If the trial succeeds, connect those nodes
    - “If the coin lands heads, connect those nodes”
- Repeat for all N(N-1)/2 pairs

Examples

These 3 networks were generated with N=100 and p=0.03; nodes at the bottom ended up isolated

The binomial distribution

- The distribution of the probability of obtaining x successes in N independent trials, in which each trial has probability of succeeding p

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

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

Properties

- Expected number of links

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

- Average degree

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

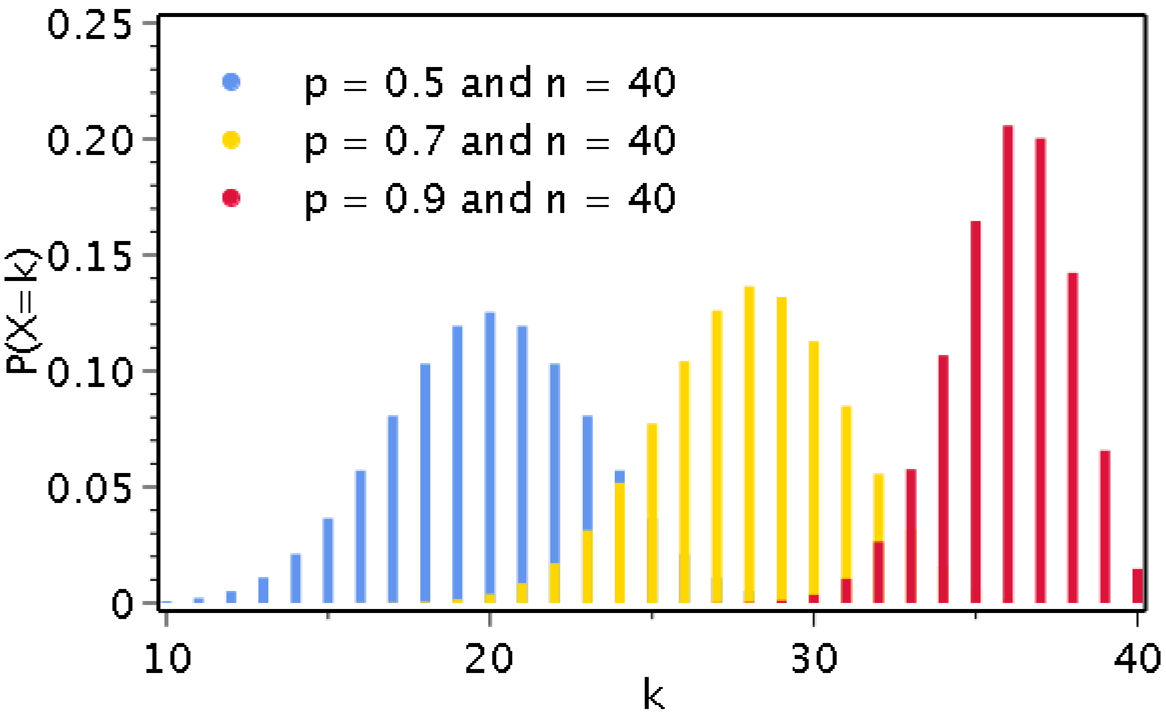
Degree distribution

- The degree distribution is simply a Binomial distribution
- Note that the maximum number of “successes” (links) of a node is N-1, hence:

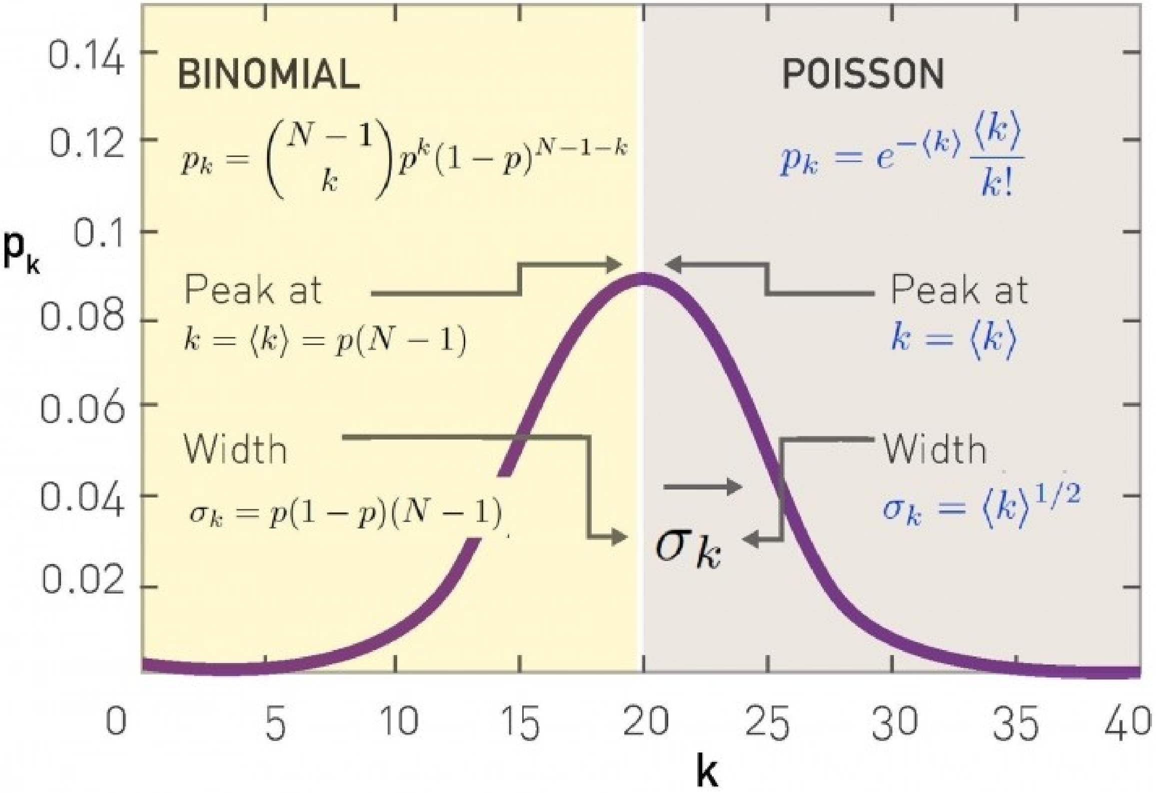
$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

Degree distribution examples

- The peak is always at <k>



Approximation with a Poisson distribution for $\langle k \rangle \ll N$



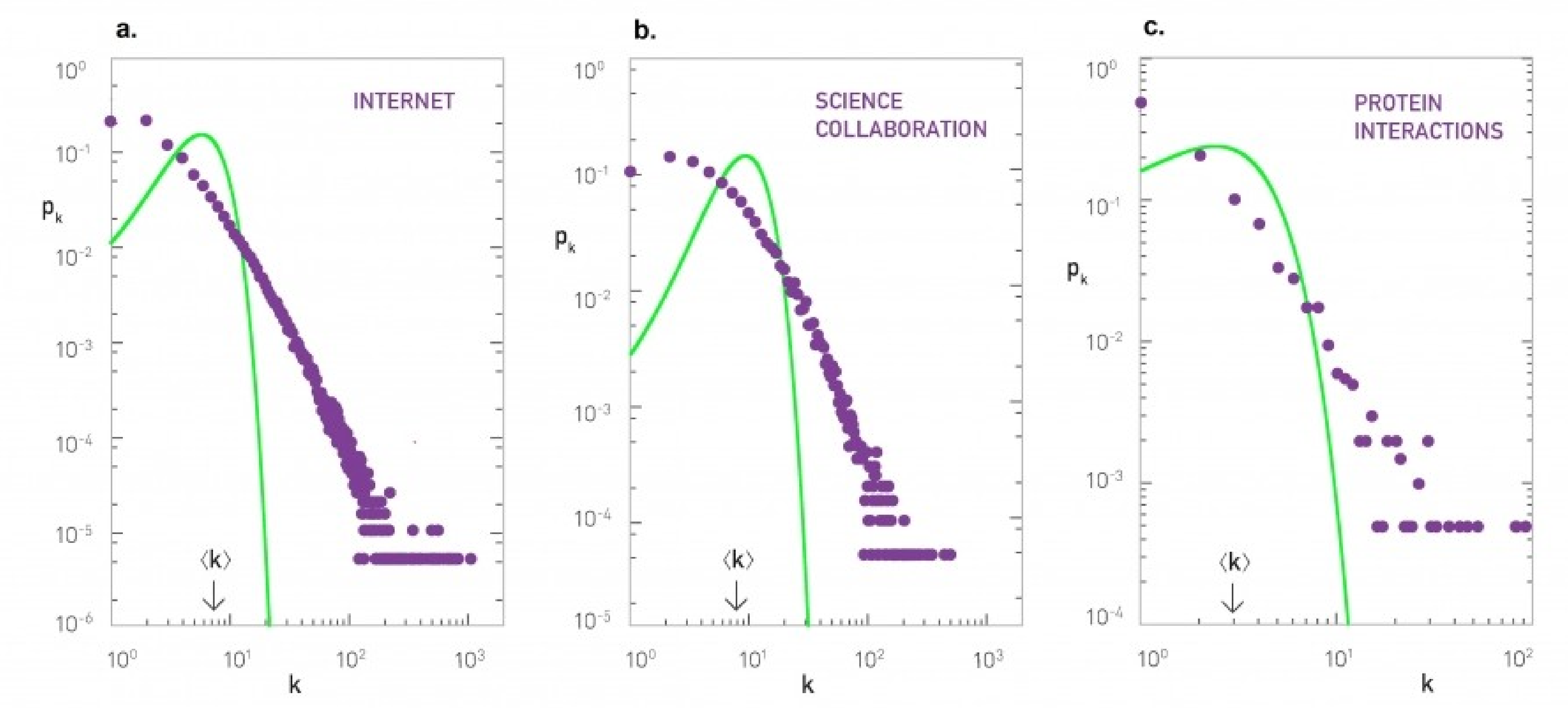
“Back of the envelope” calculations

- Suppose $N = 7 \times 10^9$
- Suppose $\langle k \rangle = 1{,}000$
  - A person knows the name of approx. 1,000 others
- Then on expectation $k_{max} = 1{,}185$
- $\langle k \rangle \pm \sigma$ is the range from 968 to 1,032
- Is this realistic?

Survey: how many WhatsApp contacts do you have?

https://goo.gl/forms/ovVvdnlWmZgMWdiL2

Real networks (green = $e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$)

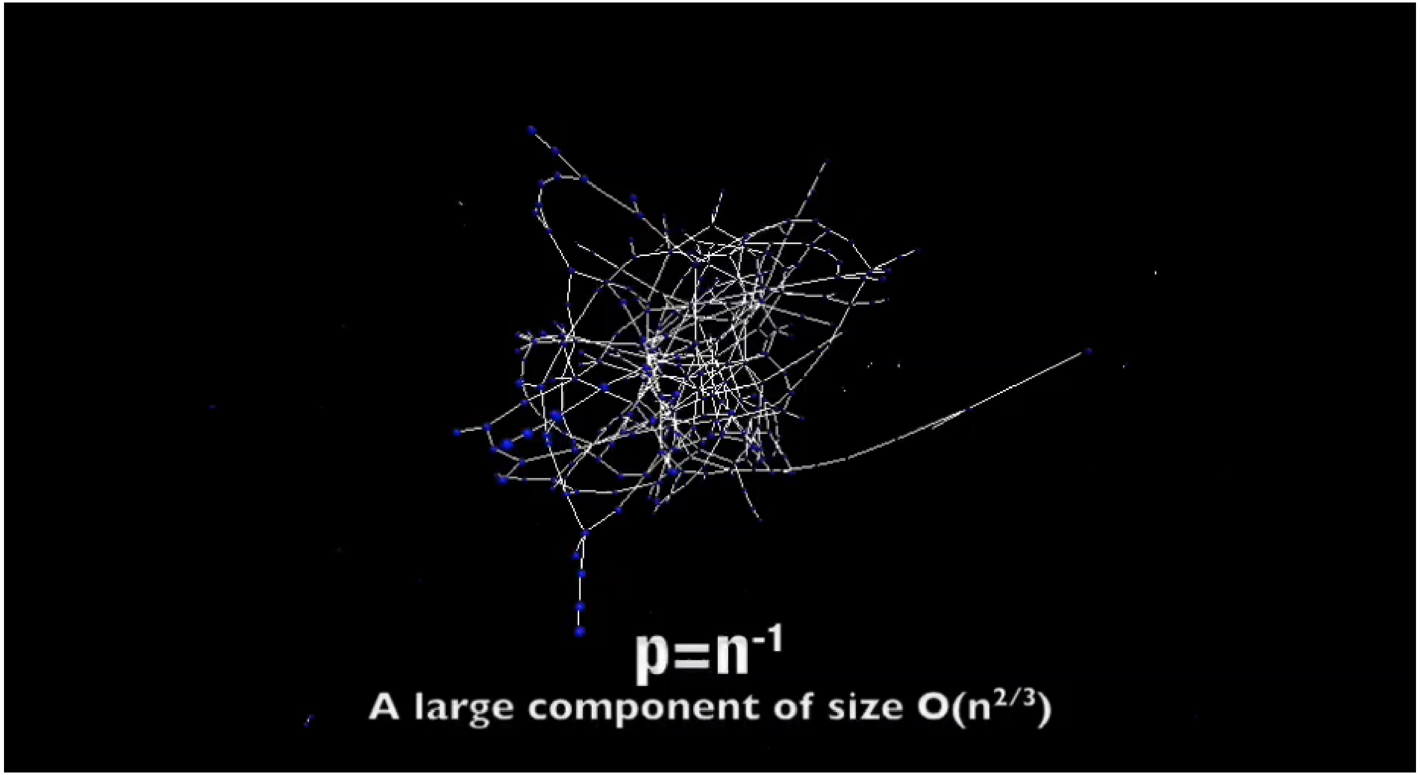


ER network as <k> increases

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

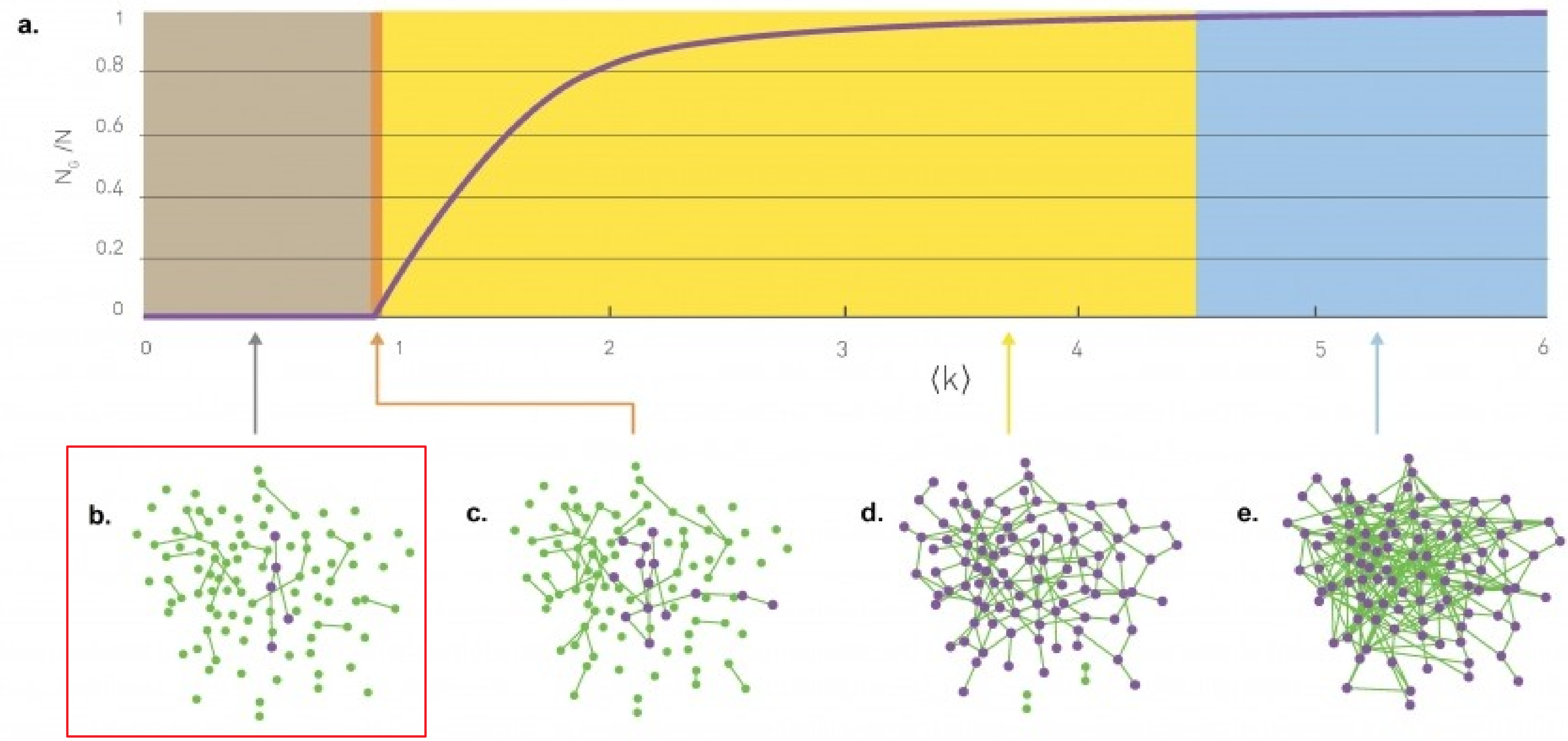
It's kind of obvious that to have a giant connected it is necessary that <k> = 1, ER proved it's sufficient in 1959

Visualization of increasing p

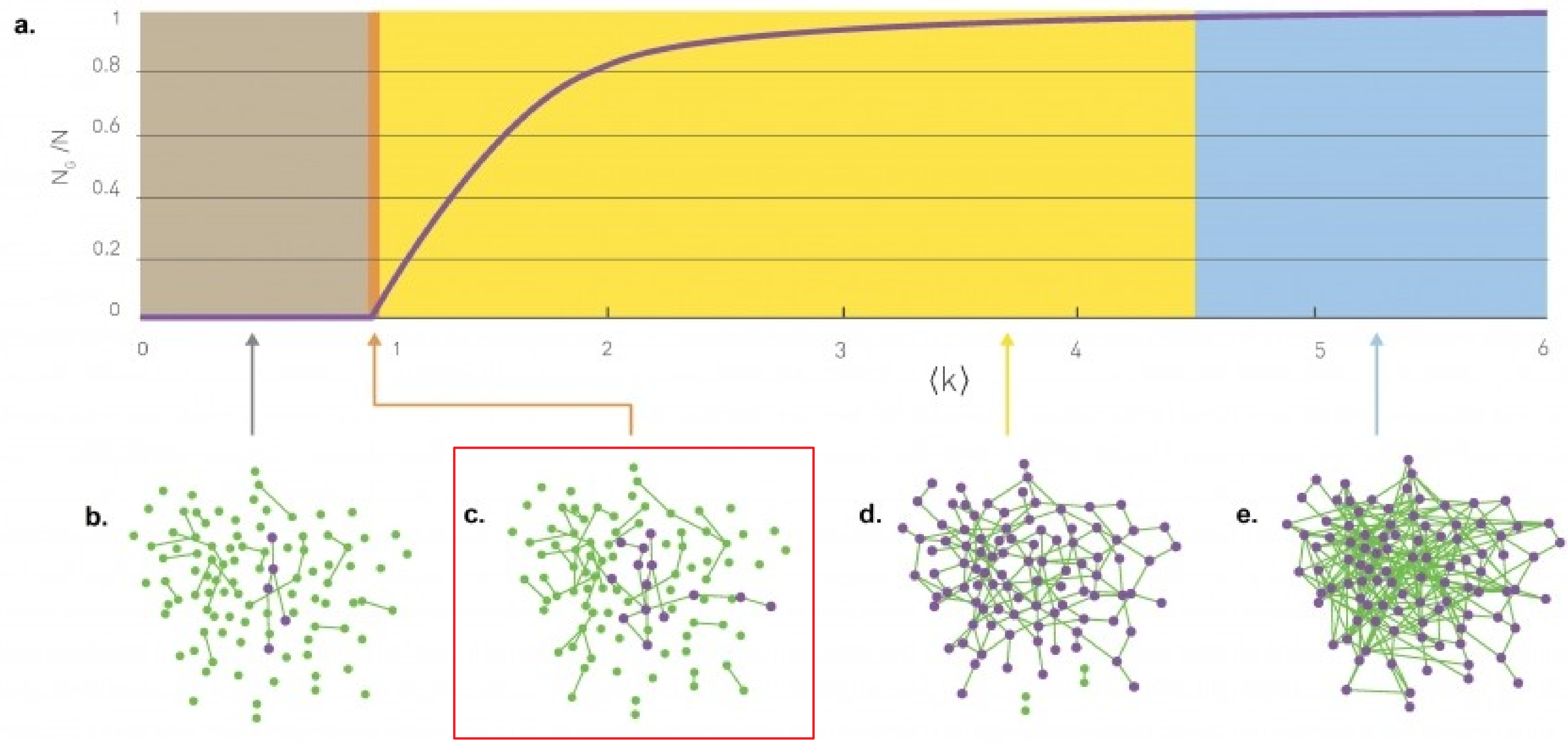


http://networksciencebook.com/images/ch-03/video-3-2.m4v

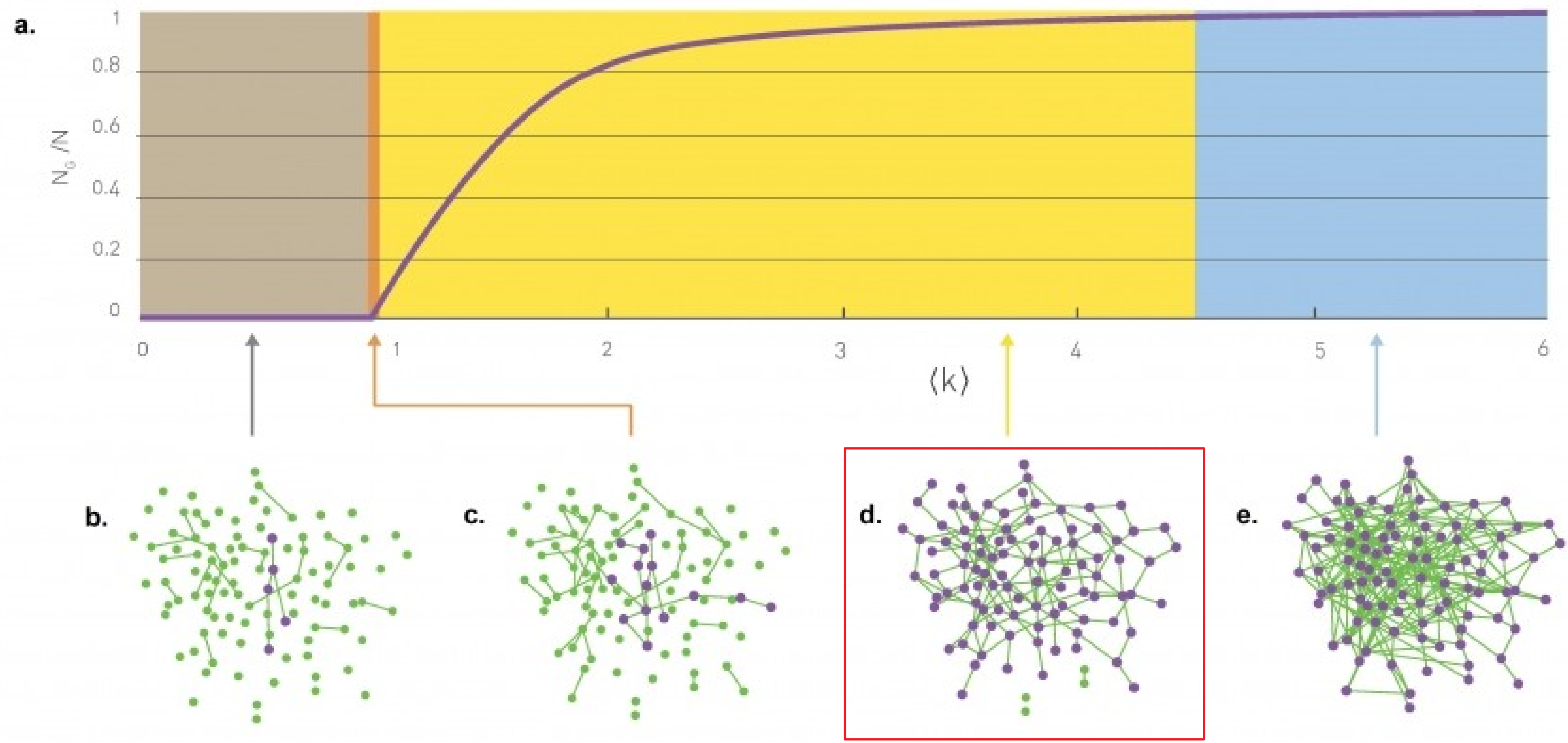
Sub-critical regime: $\langle k \rangle < 1$



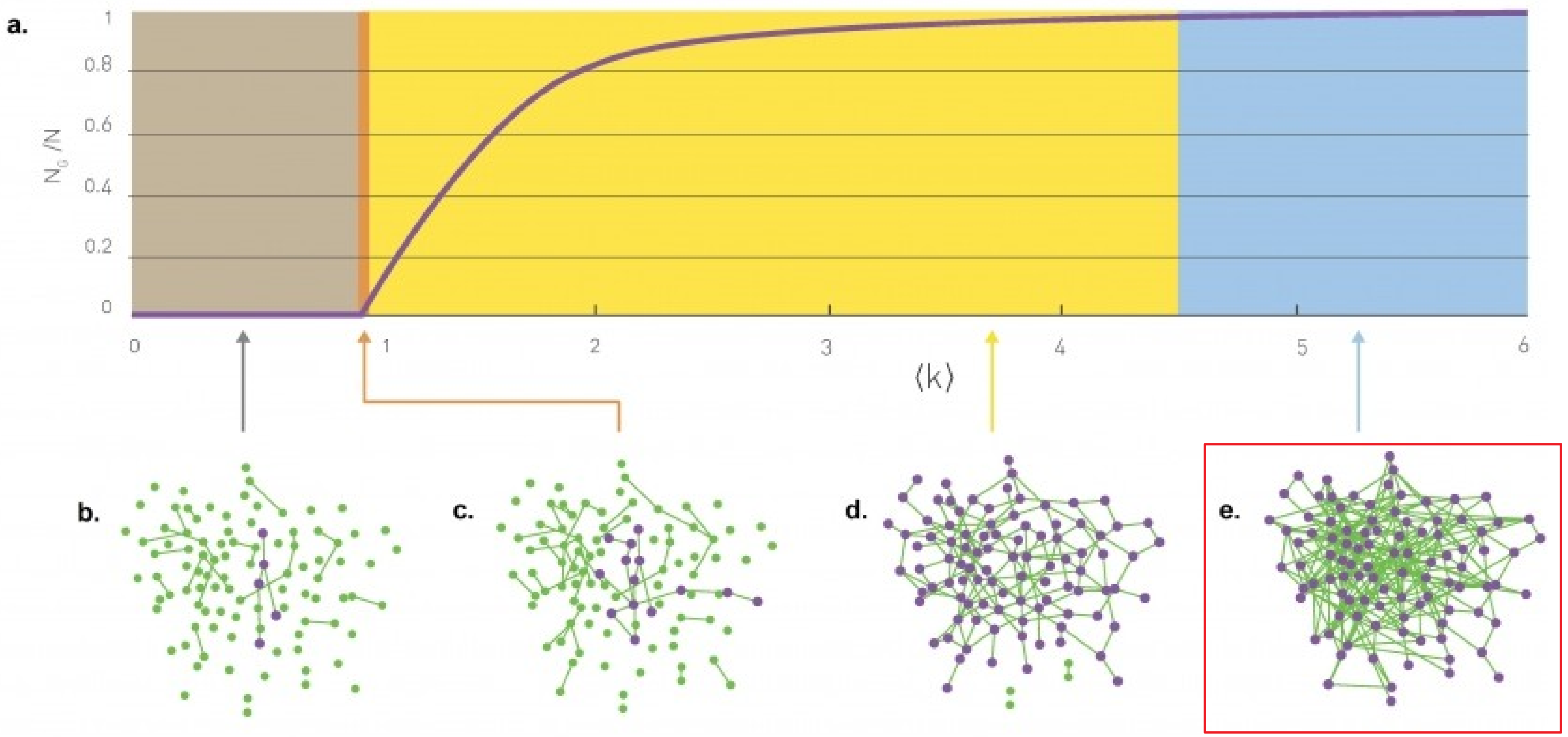
Critical point: $\langle k \rangle = 1$



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



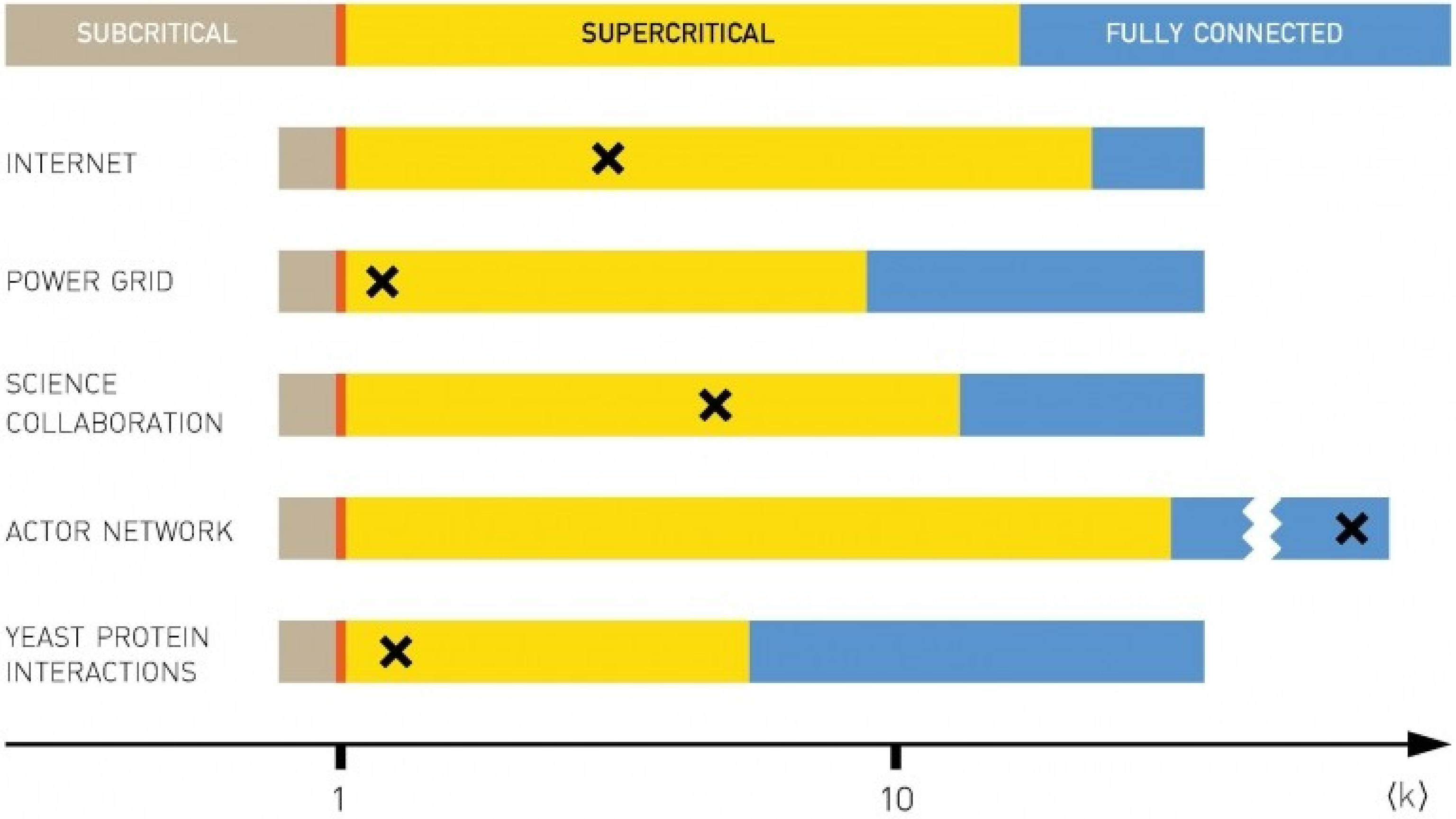
Connected regime: $\langle k \rangle > \log N$



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

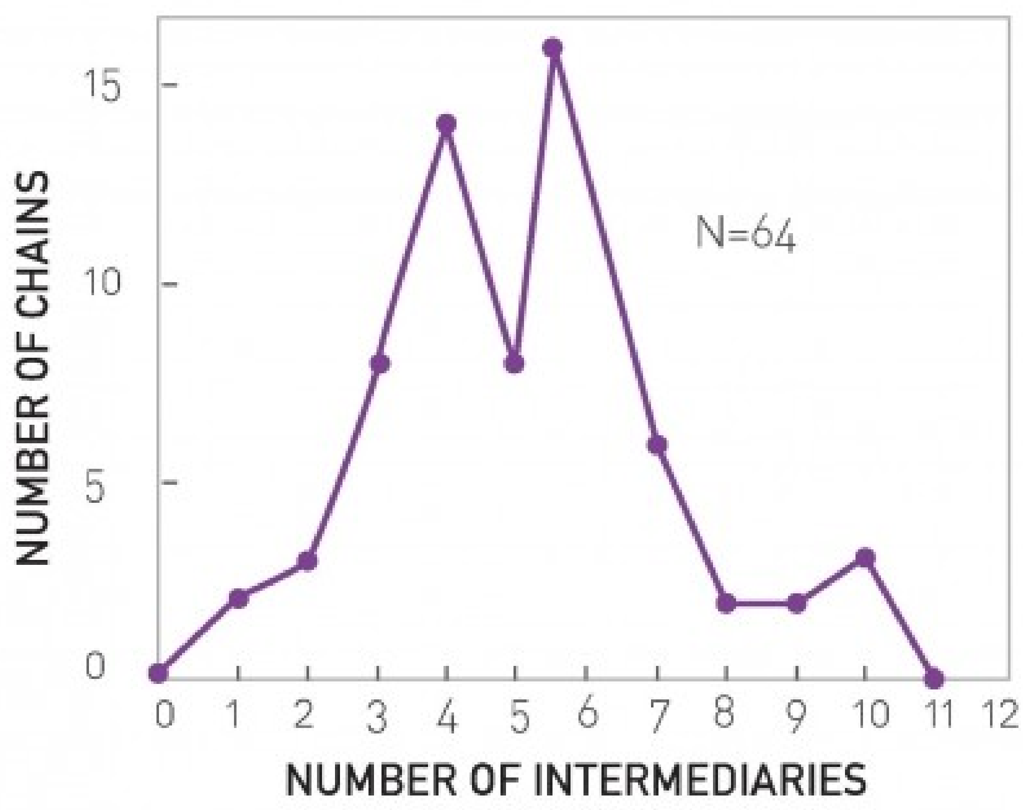
Most real networks are supercritical: $\langle k \rangle > 1$

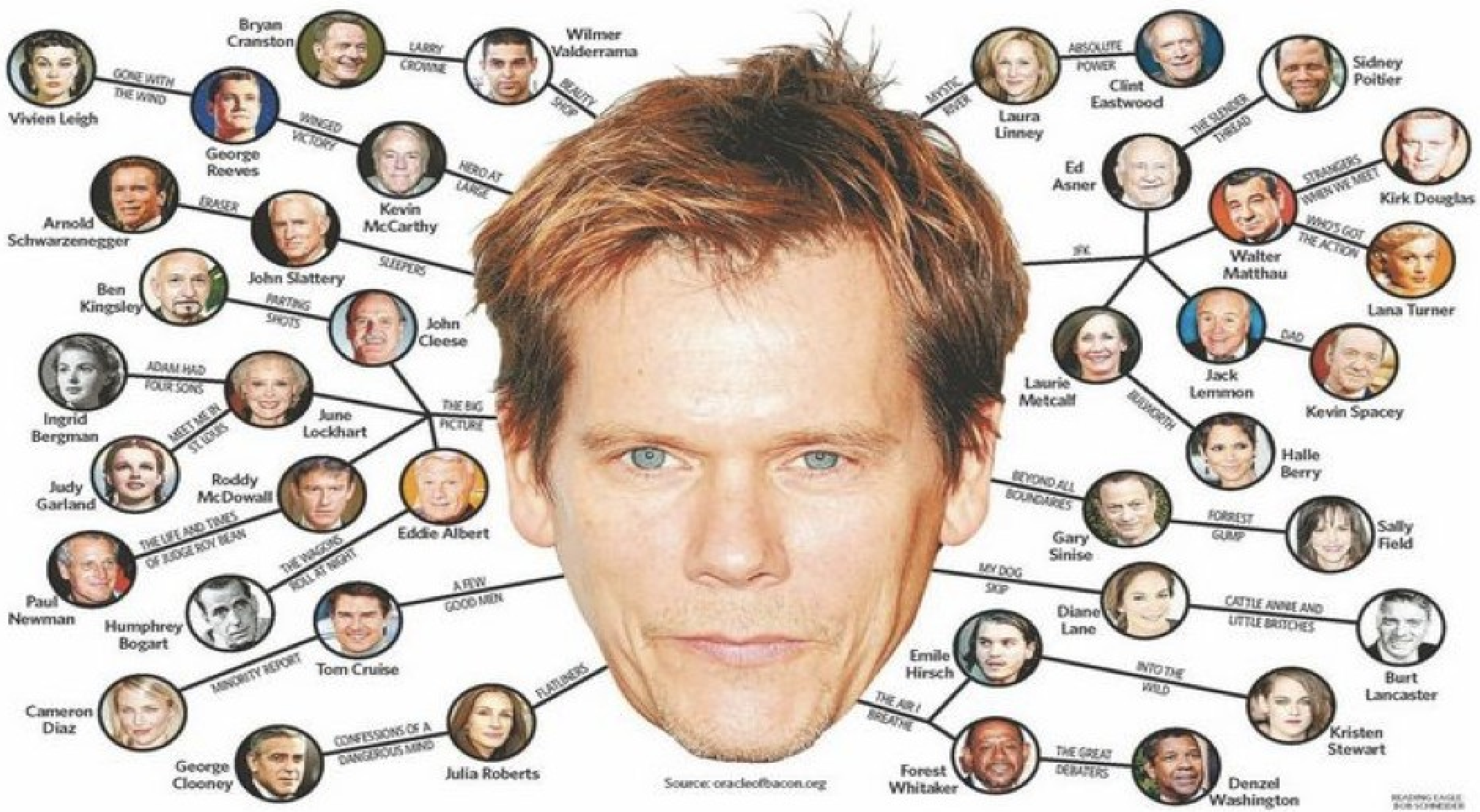


Small-world phenomenon a.k.a. “six degrees of separation”

Milgram’s experiment in 1967

- Targets: (1) a stock broker in Boston, MA and (2) a student in Sharon, MA
- Sources: residents of Wichita and Omaha
- Materials: a short summary of the study’s purpose, a photograph, the name, address and information about the target person
- Request: to forward the letter to a friend, relative or acquaintance who is most likely to know the target person.
- **<u>64 of 296</u> letters reached destination**





Vivien Leigh
GONE WITH THE WIND
George Reeves
Bryan Cranston
LARRY CROWNE
Wilmer Valderrama
BEAUTY SHOP
WINGED VICTORY
Kevin McCarthy
HERO AT LARGE
Arnold Schwarzenegger
ERASER
John Slattery
SLEEPERS
Ben Kingsley
PARTING SHOTS
John Cleese
Ingrid Bergman
ADAM HAD FOUR SONS
June Lockhart
MEET ME IN ST. LOUIS
Judy Garland
THE BIG PICTURE
Roddy McDowall
Eddie Albert
THE LIFE AND TIMES OF JUDGE ROY BEAN
Paul Newman
THE WAGONS ROLL AT NIGHT
Humphrey Bogart
A FEW GOOD MEN
Tom Cruise
MINORITY REPORT
Cameron Diaz
George Clooney
CONFESSIONS OF A DANGEROUS MIND
Julia Roberts
FLATLINERS
Source: oracleofbacon.org
MYSTIC RIVER
Laura Linney
ABSOLUTE POWER
Clint Eastwood
Sidney Poitier
THE SLENDER THREAD
Ed Asner
STRANGERS WHEN WE MEET
Kirk Douglas
JFK
Walter Matthau
WHO'S GOT THE ACTION
Lana Turner
Laurie Metcalf
Jack Lemmon
DAD
Kevin Spacey
BLOWORTH
Halle Berry
BEYOND ALL BOUNDARIES
Gary Sinise
FORREST GUMP
Sally Field
MY DOG SKIP
Diane Lane
CATTLE ANNIE AND LITTLE BRITCHES
Burt Lancaster
Emile Hirsch
INTO THE WILD
Kristen Stewart
THE AIR I BREATHE
Forest Whitaker
THE GREAT DEBATERS
Denzel Washington

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 (around six at most)
- Formally
  - The distance between two randomly chosen nodes in a network is short

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

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

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

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

Empirical average and maximum distances

Network	N	L	‹k›	‹d›	d_{max}	$\ln N/\ln \langle k \rangle$
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:

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

Clustering coefficient

or

”a friend of a friend is my friend”

Clustering coefficient C_i of node i

- $C_i = 0 \Rightarrow$ 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 $C_i = \frac{2\langle L_i \rangle}{k_i(k_i-1)}$

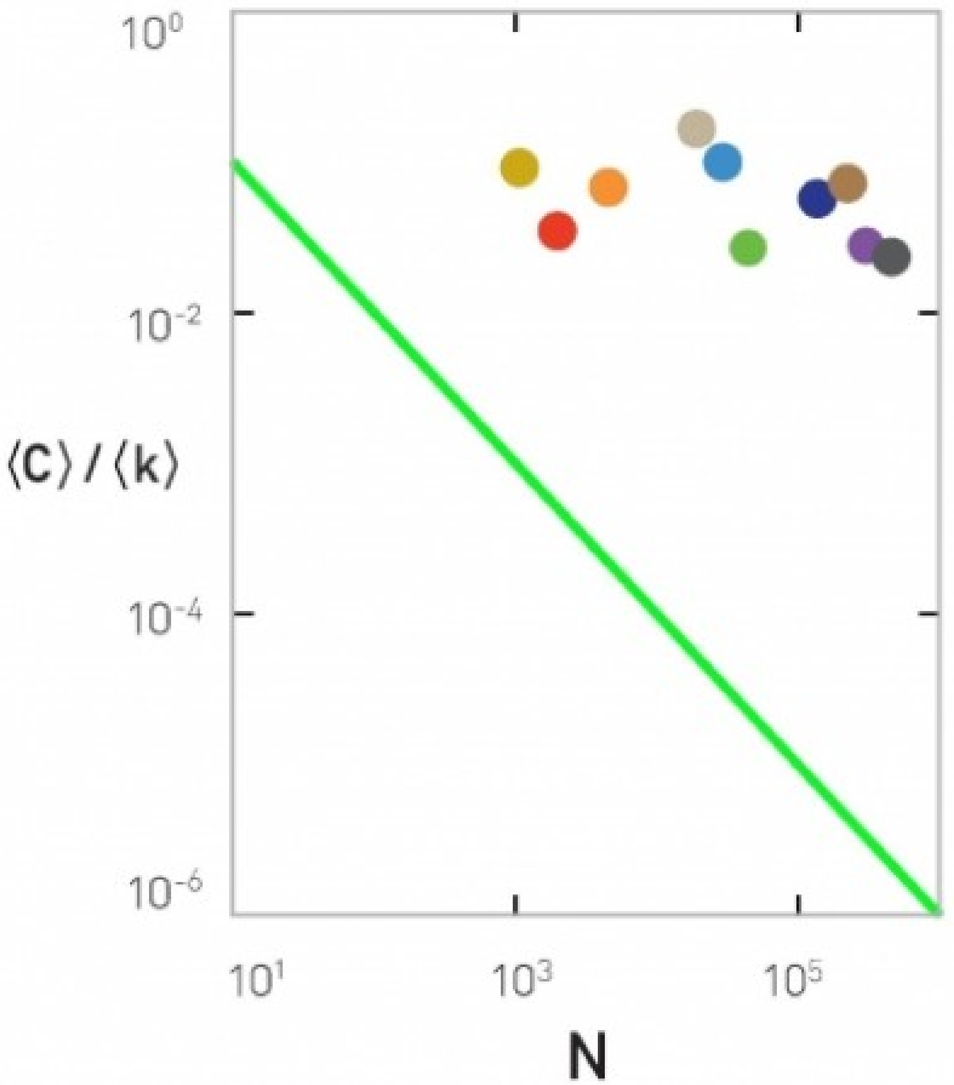
$$= \frac{2p\frac{k_i(k_i-1)}{2}}{k_i(k_i-1)} = \frac{\langle k \rangle}{N}$$

In an ER graph

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

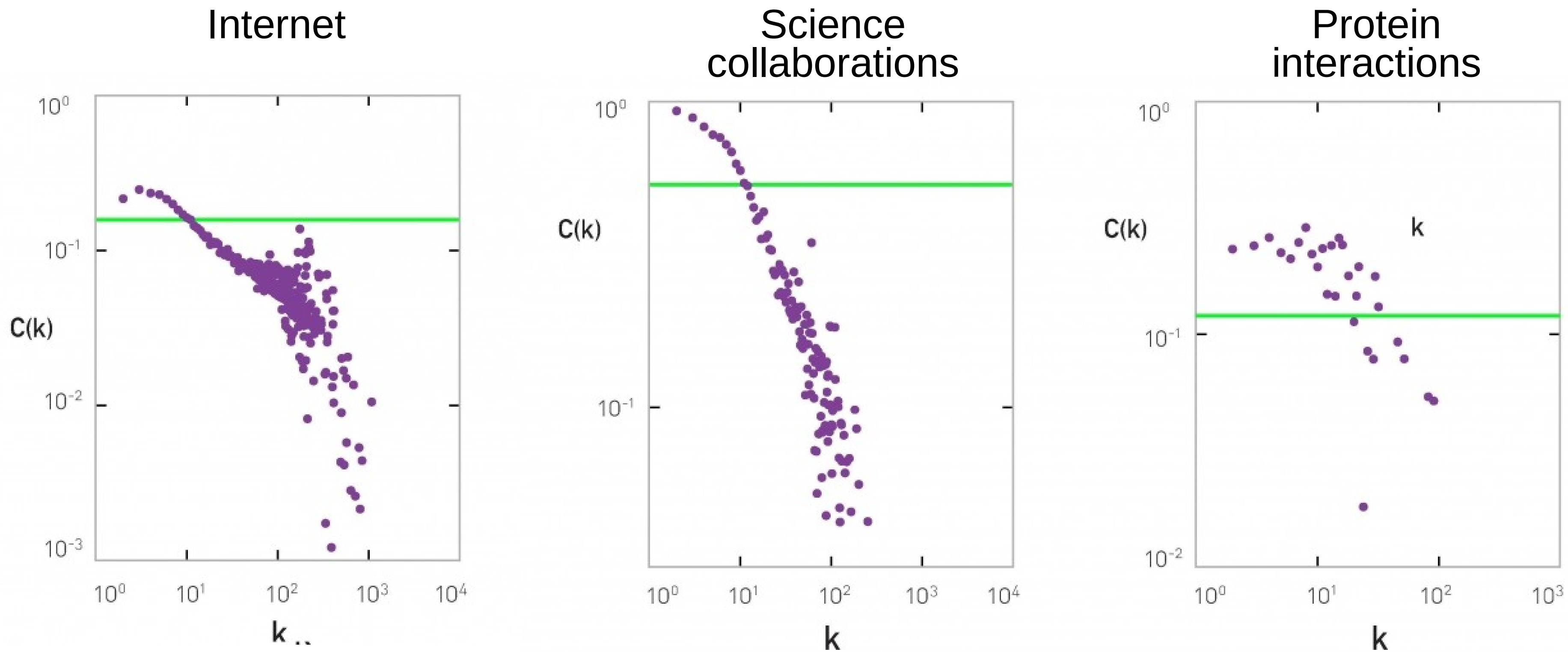
If <k> is fixed, large networks should have smaller clustering coefficient

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



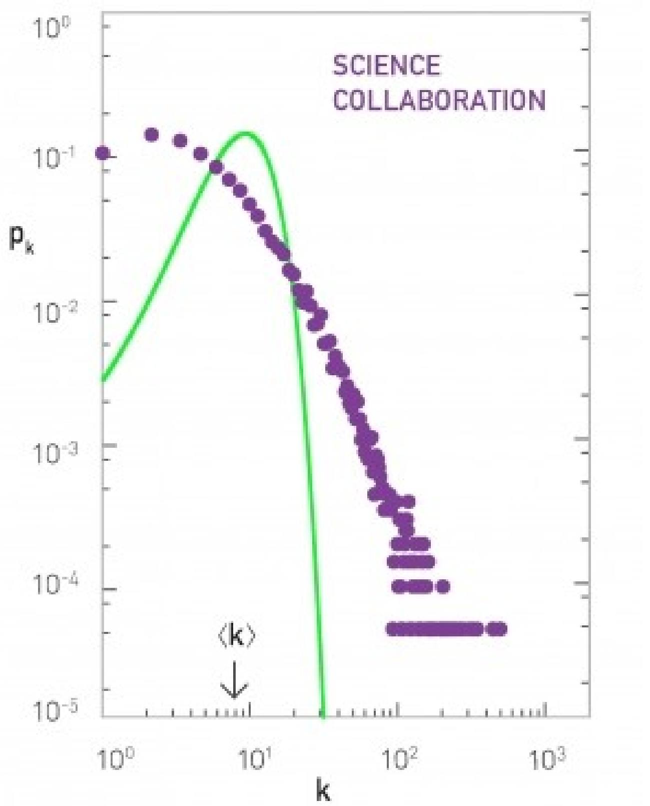
To re-cap …

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

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

- Observed

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

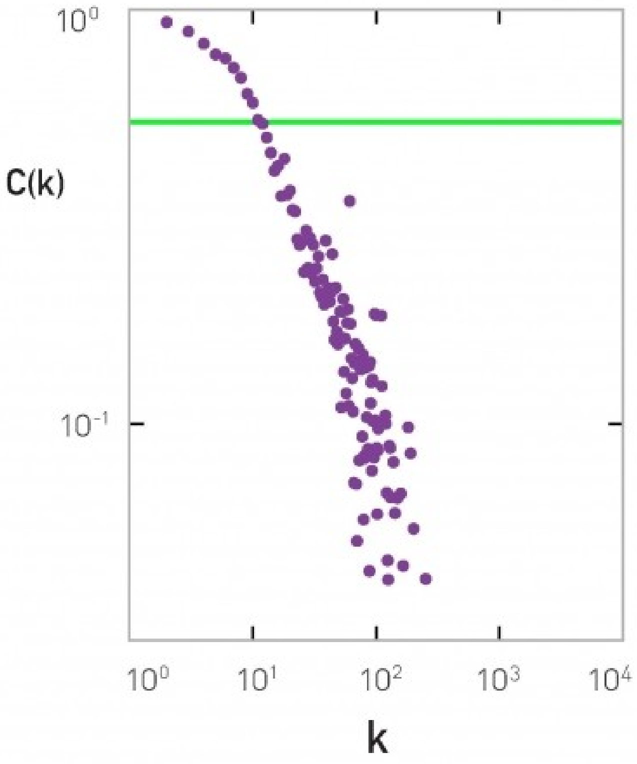
‹d›	d_{max}	ln$N$/ln‹k›
6.98	26	6.58
11.27	93	8.31
18.99	46	8.66
11.72	39	11.42
5.88	18	18.4
5.35	15	4.81
3.91	14	3.04
11.21	42	5.55
2.98	8	4.04
5.61	14	7.14

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

- Predicted

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

- Observed
  Clustering coefficient depends on the degree of the node



Why do we study the ER model?

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

Exercise [B. 2016, Ex. 3.11.1]

- Consider an ER graph with $N=3{,}000$ $p=10^{-3}$
  1) What is the expected number of links $\langle L \rangle$?
  2) In which regime is the network?
  3) What is N_{cr} so that there is only one component?
  4) What it the average degree if the network has N_{cr} nodes?
  5) What is the expected distance $\langle d \rangle$?