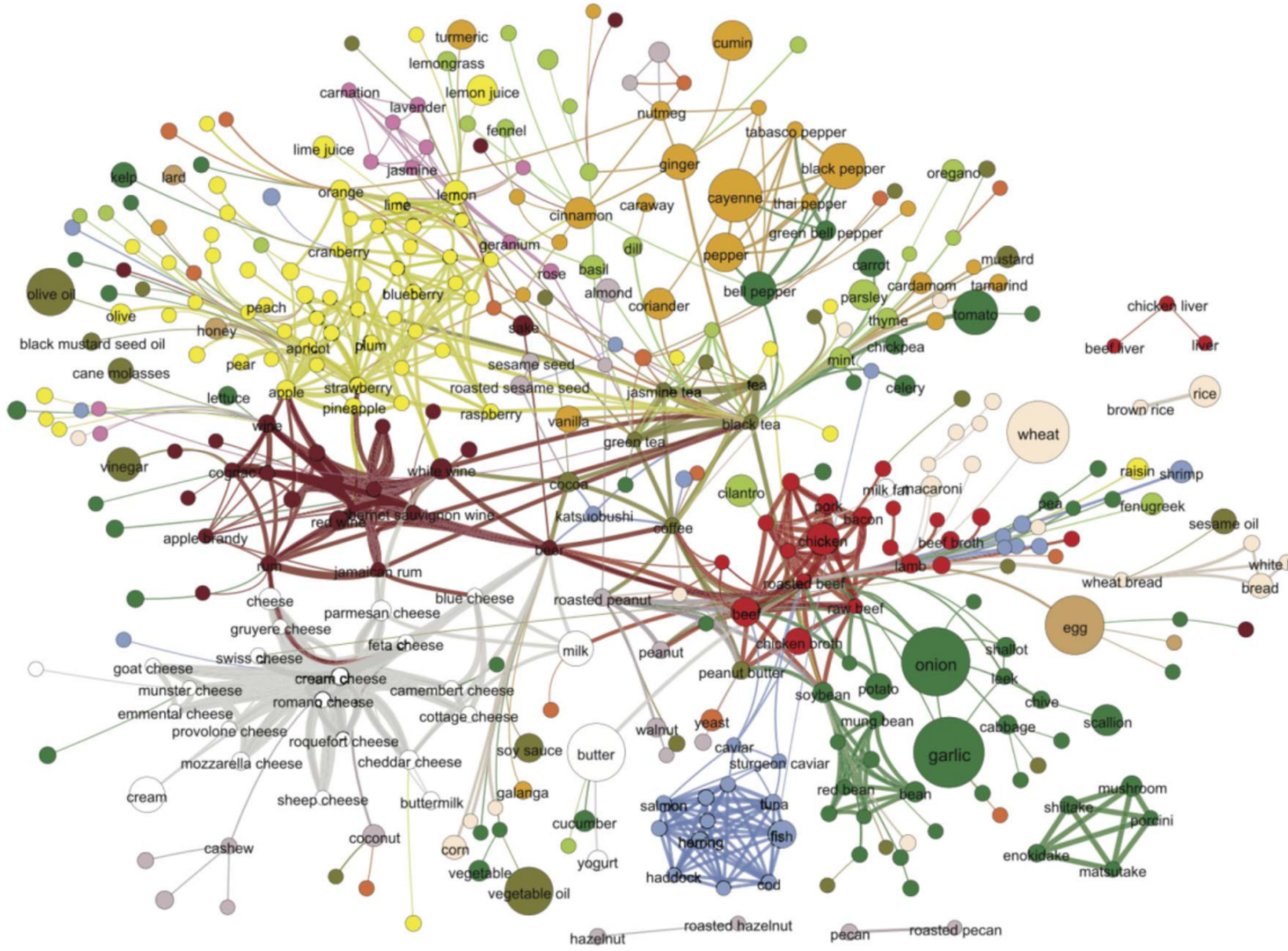
Social Network Analysis Lecture 13

2162-F23

2023-03-20



Flavor Network

Ahn, Yong-Yeol, et al. "Flavor network and the principles of food pairing." Scientific reports 1 (2011): 196.

What do we do today?

Distributions and models

→ Statistical distribution: Definition, Calculation, Interpretation, Degree distribution

→ Random network model

→ Stanley Milgram (Small-world) Shortest path in the real world

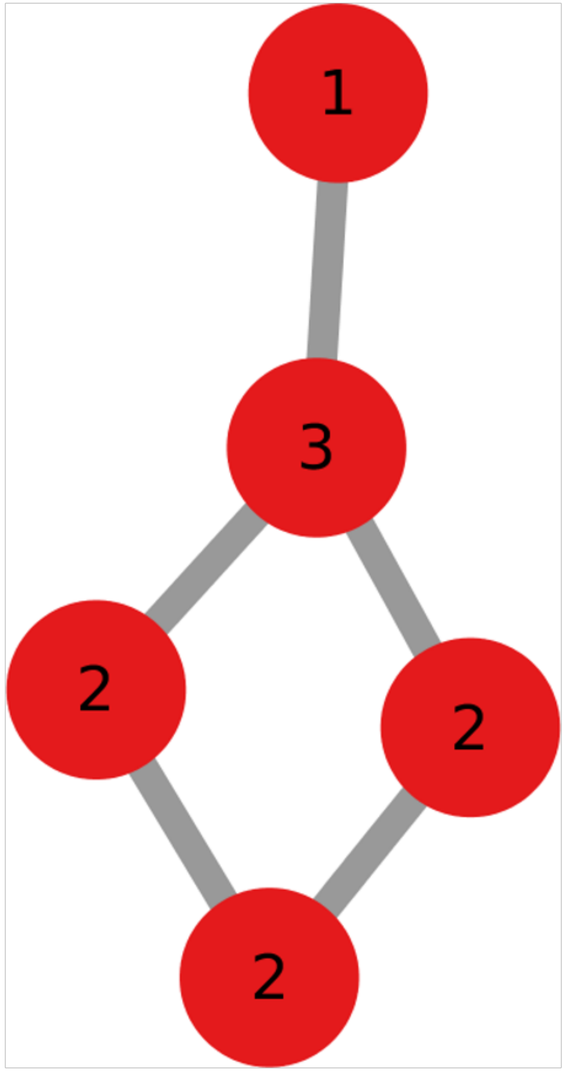
Degree distribution

Degree

Number of node connections k

Each node has one!

Any patterns?

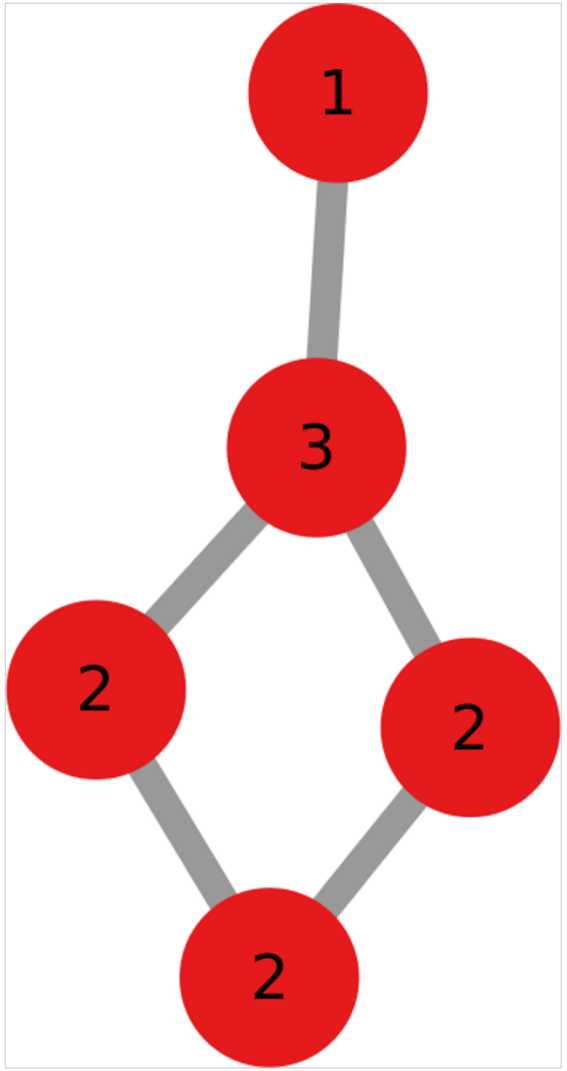


Node label = degree k_i

Degree

- On small networks it makes sense to ask which nodes or links are most important

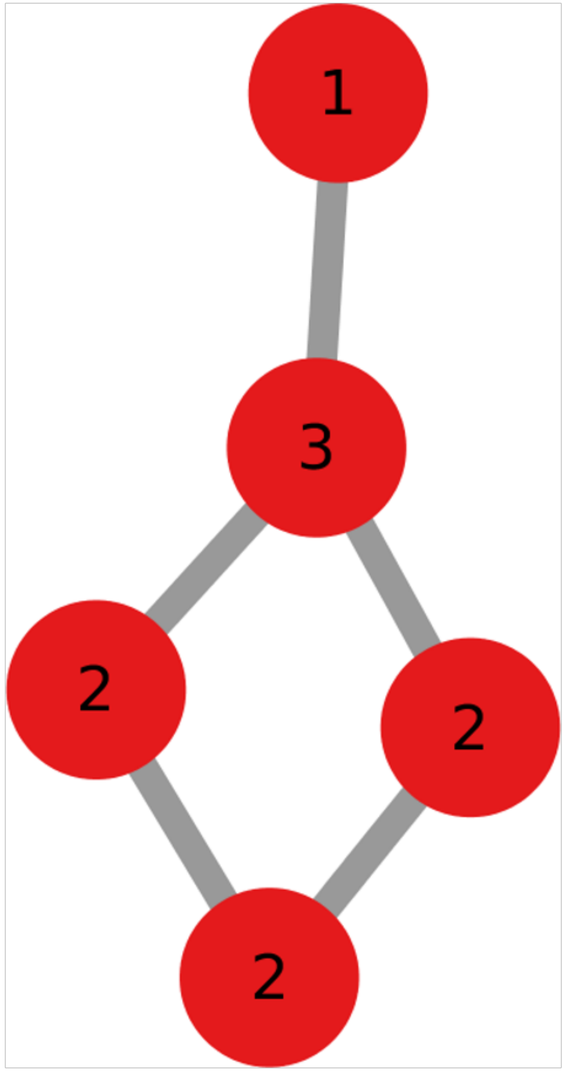
- On large networks **it does not**



Node label = degree k_i

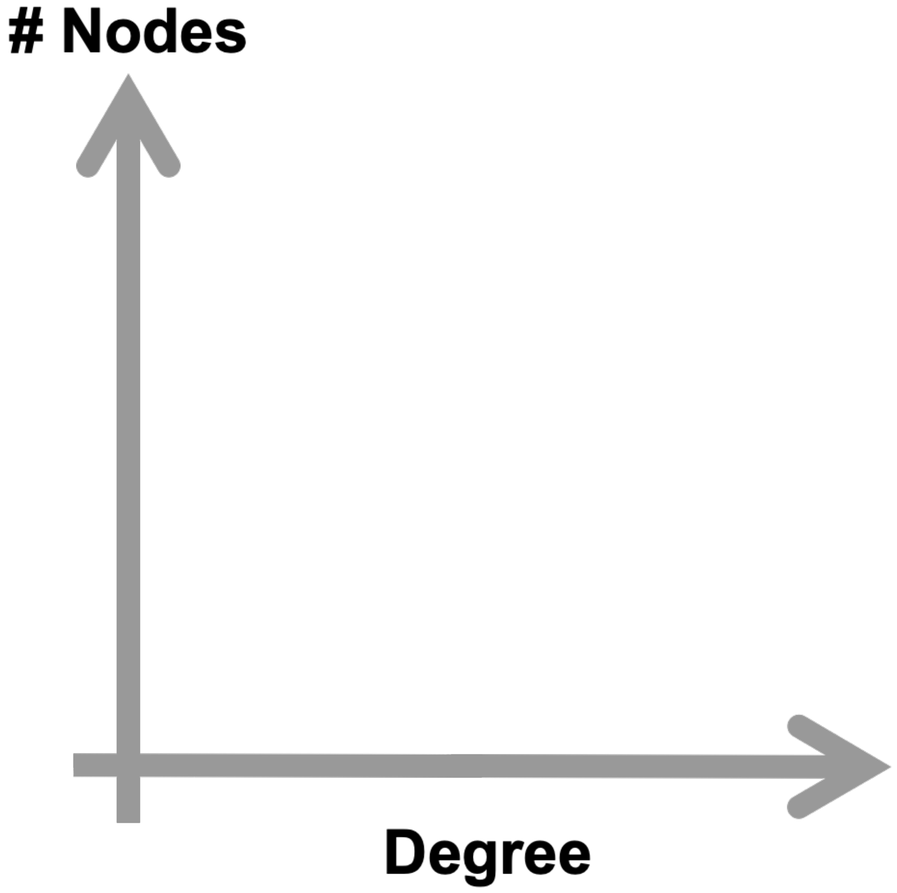
Degree

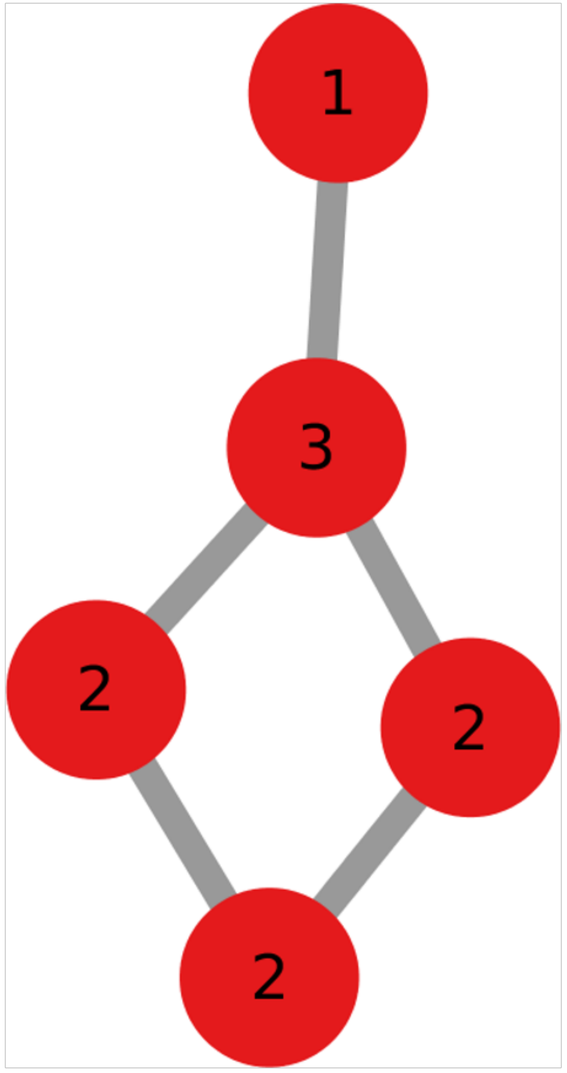
- On small networks it makes sense to ask which nodes or links are most important
- On large networks **it does not**
- **Solution:** statistical approach
- Instead of focusing on individual nodes and links, we consider **classes** of nodes and links with similar properties



Node label = degree k_i

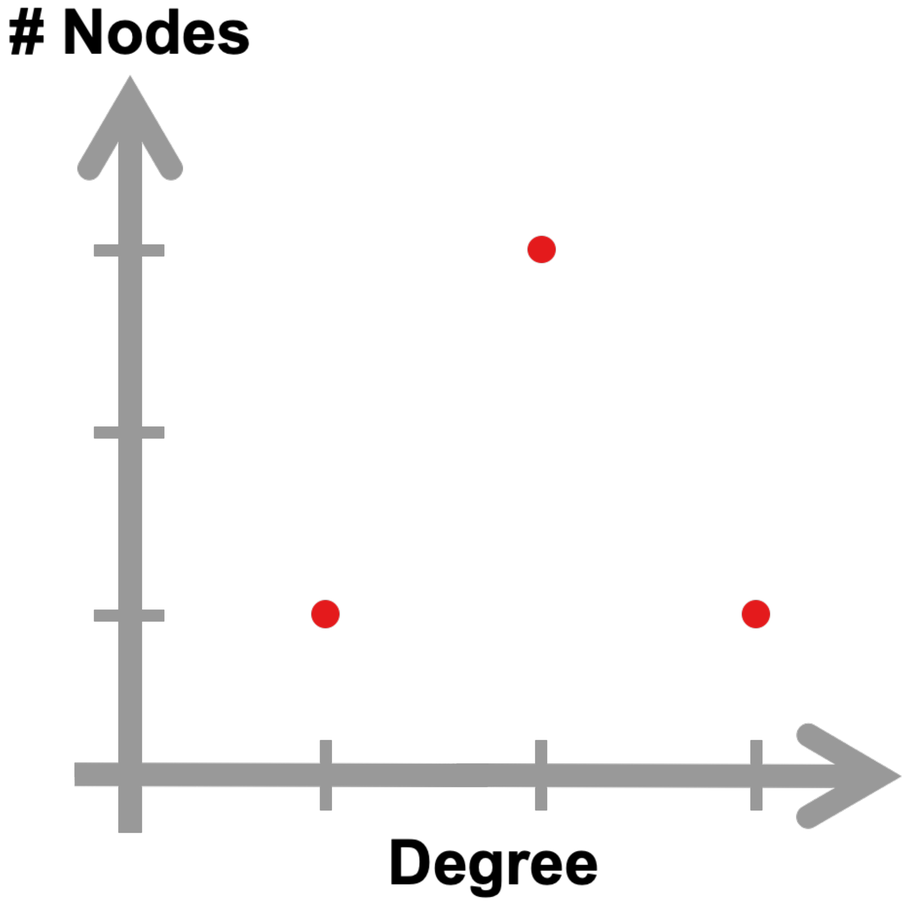
Degree distribution



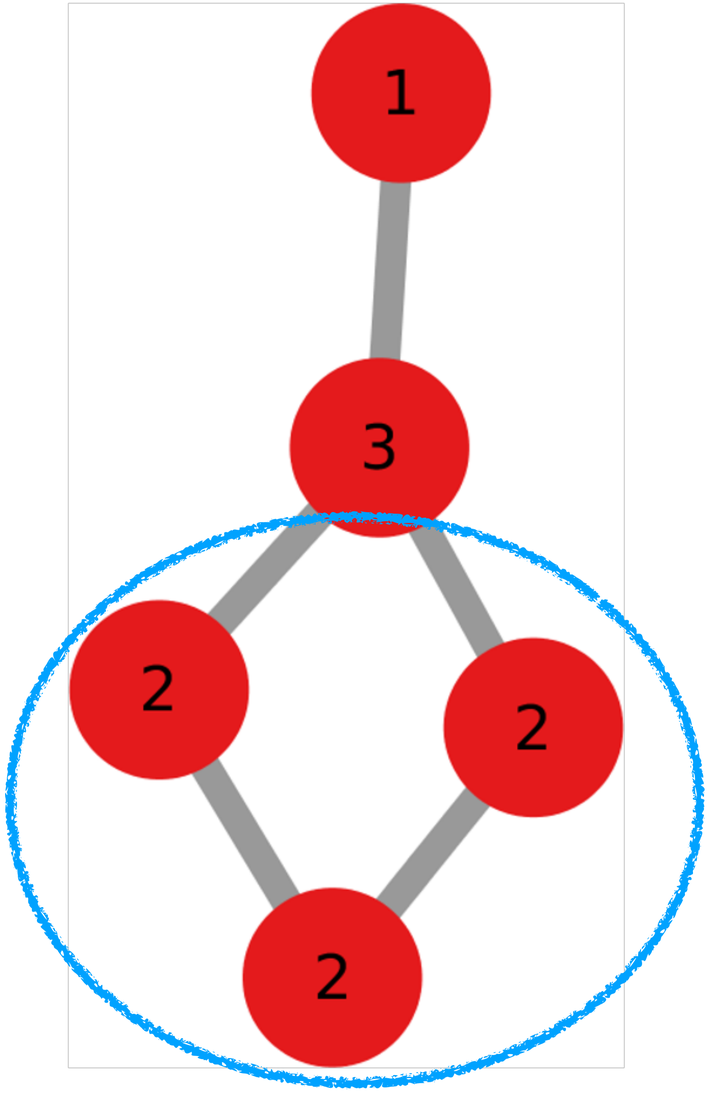


Node label = degree k_i

Degree distribution



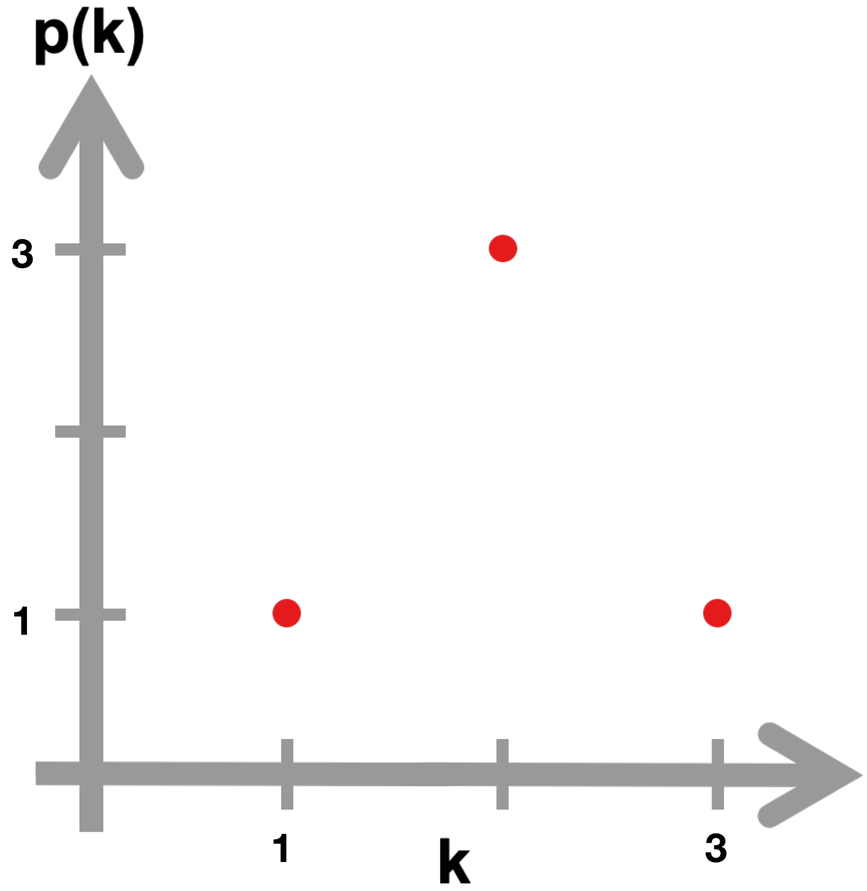
How many nodes with degree ***k* = *2***?

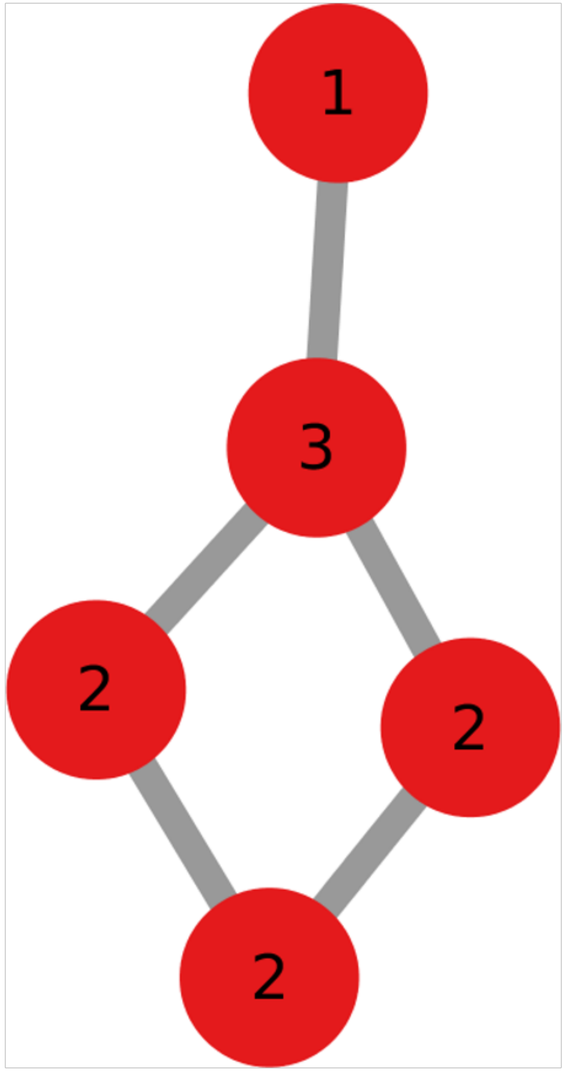


Node label = degree k_i

Degree distribution

Probability distribution: plot of probability p_k versus k

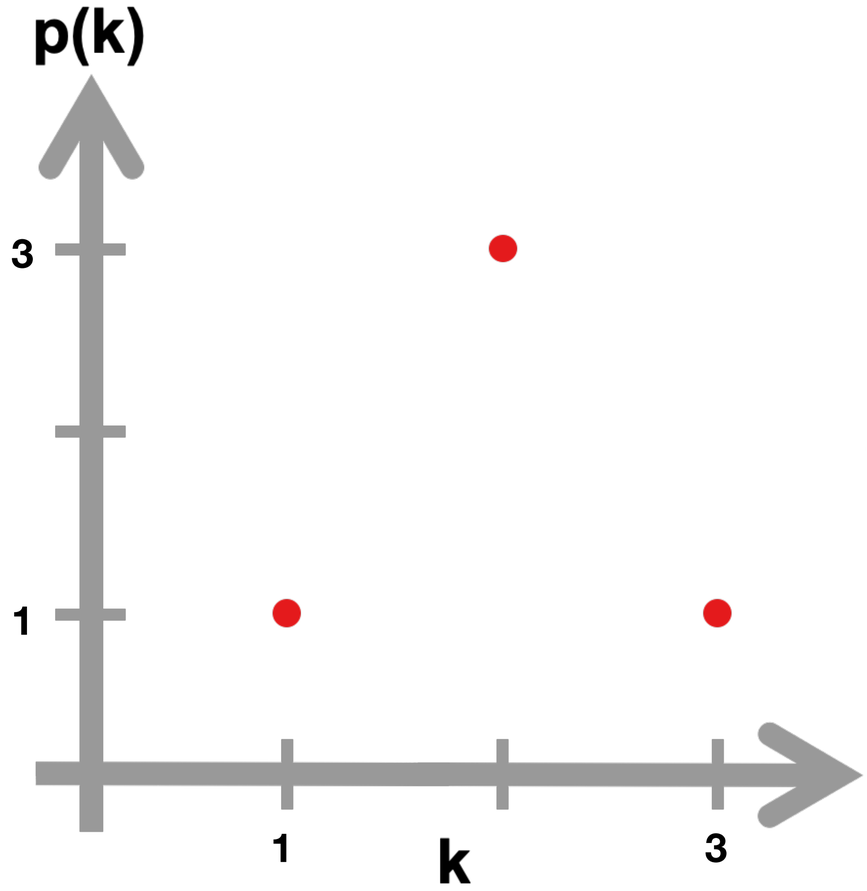




Node label = degree k_i

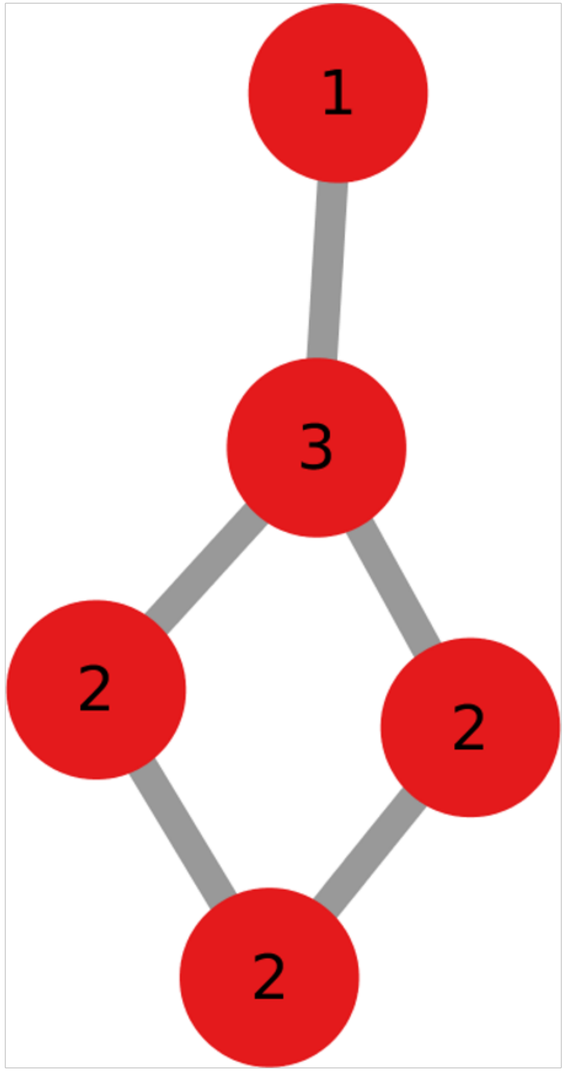
Degree distribution

Probability distribution: plot of probability p_k versus k



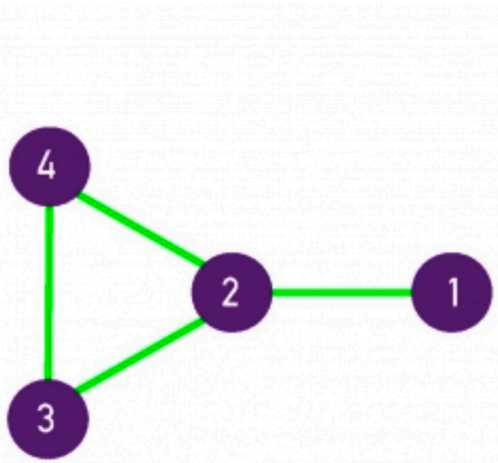
For large N, the frequency of k becomes the **probability** p_k of having degree k

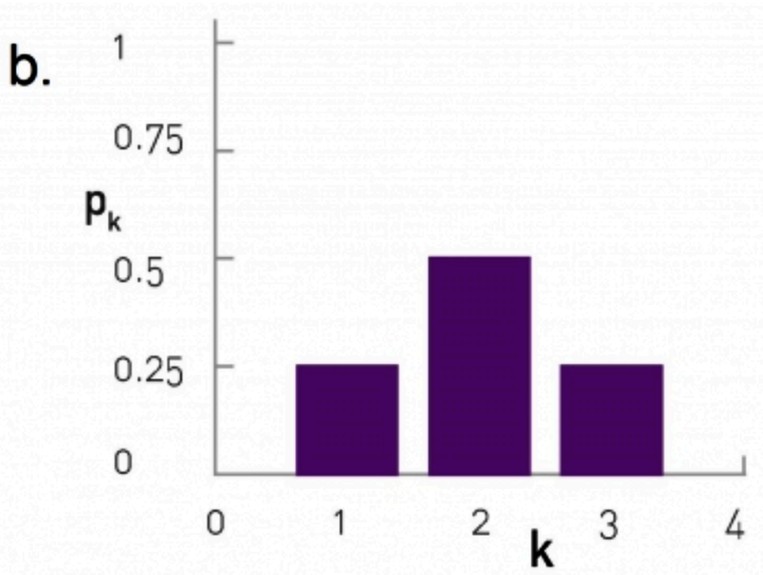
Normalise y-axis, easier to compare networks



Node label = degree k_i

Degree distribution

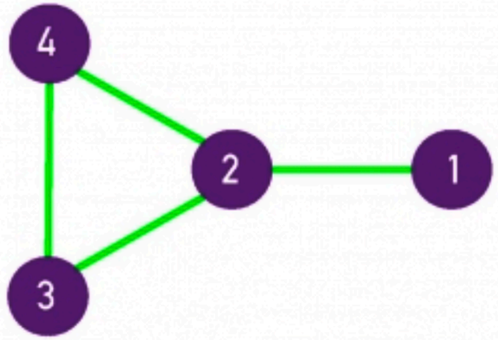


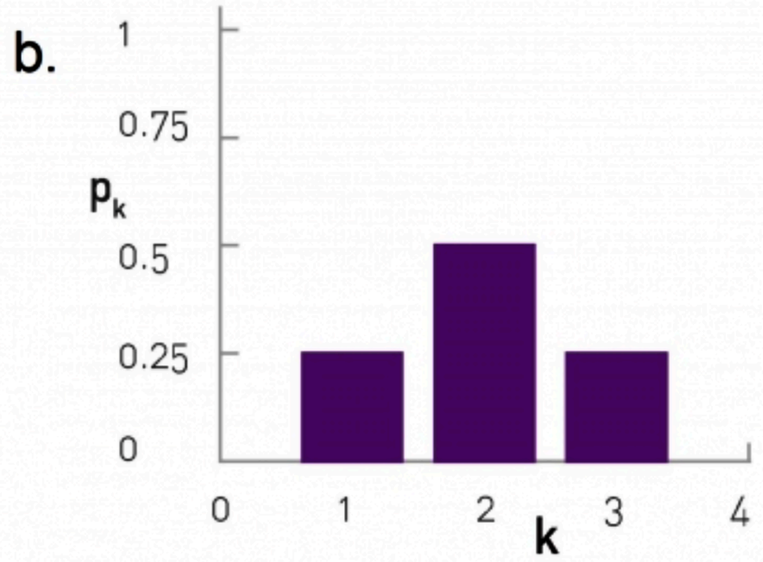


Let's make the histogram

We have:
p1 = 1/4 (1 node has $k_1 = 1$)
p2 = 1/2 (2 nodes have $k_3 = k_4 = 2$)
p3 = 1/4 (as $k_2 = 3$).

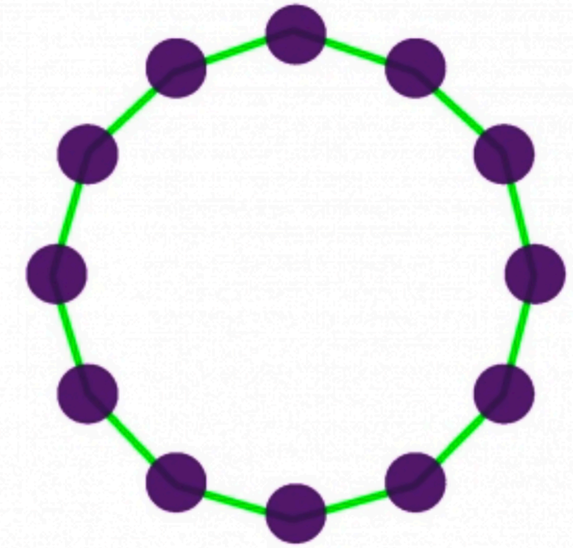
Degree distribution

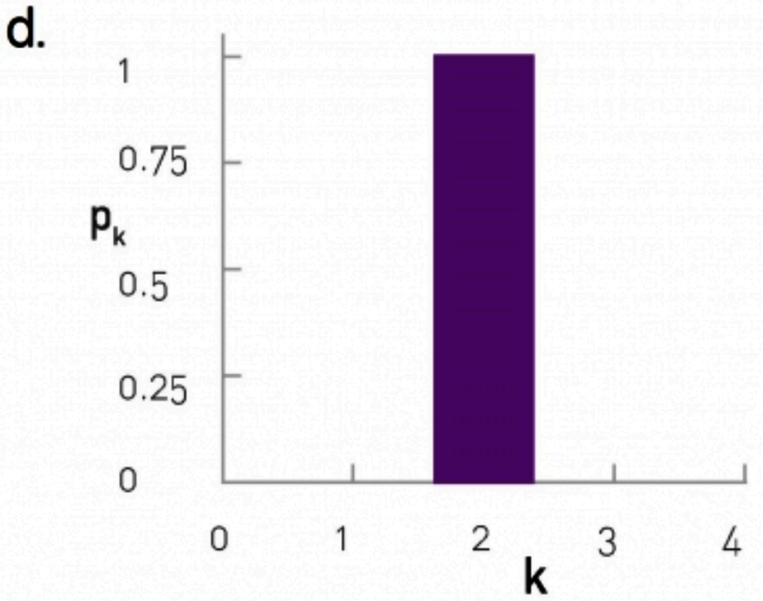




Let's make the histogram

We have:
p1 = 1/4 (1 node has $k_1 = 1$)
p2 = 1/2 (2 nodes have $k_3 = k_4 = 2$)
p3 = 1/4 (as $k_2 = 3$).

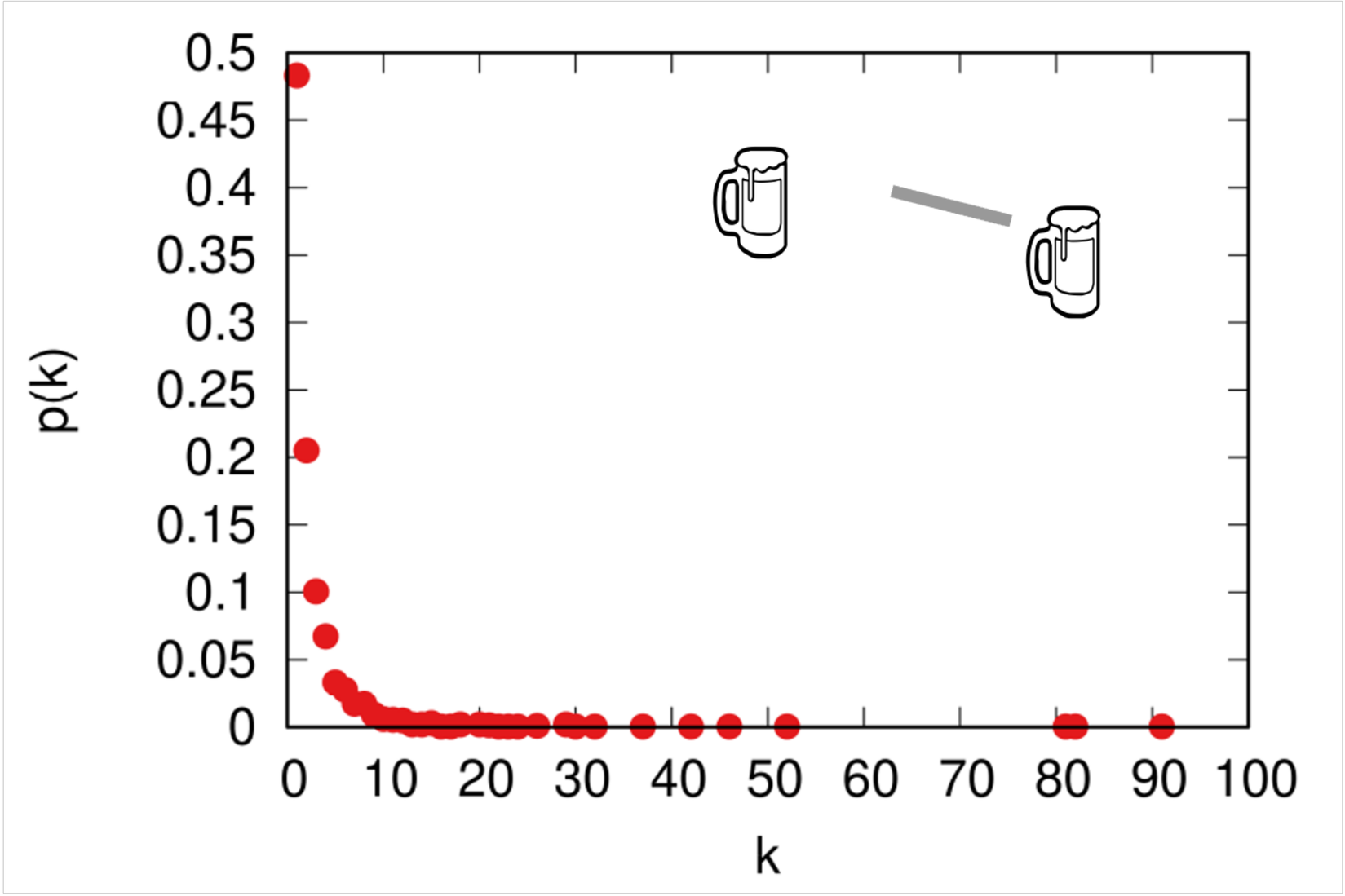




Histograms have bin size issues

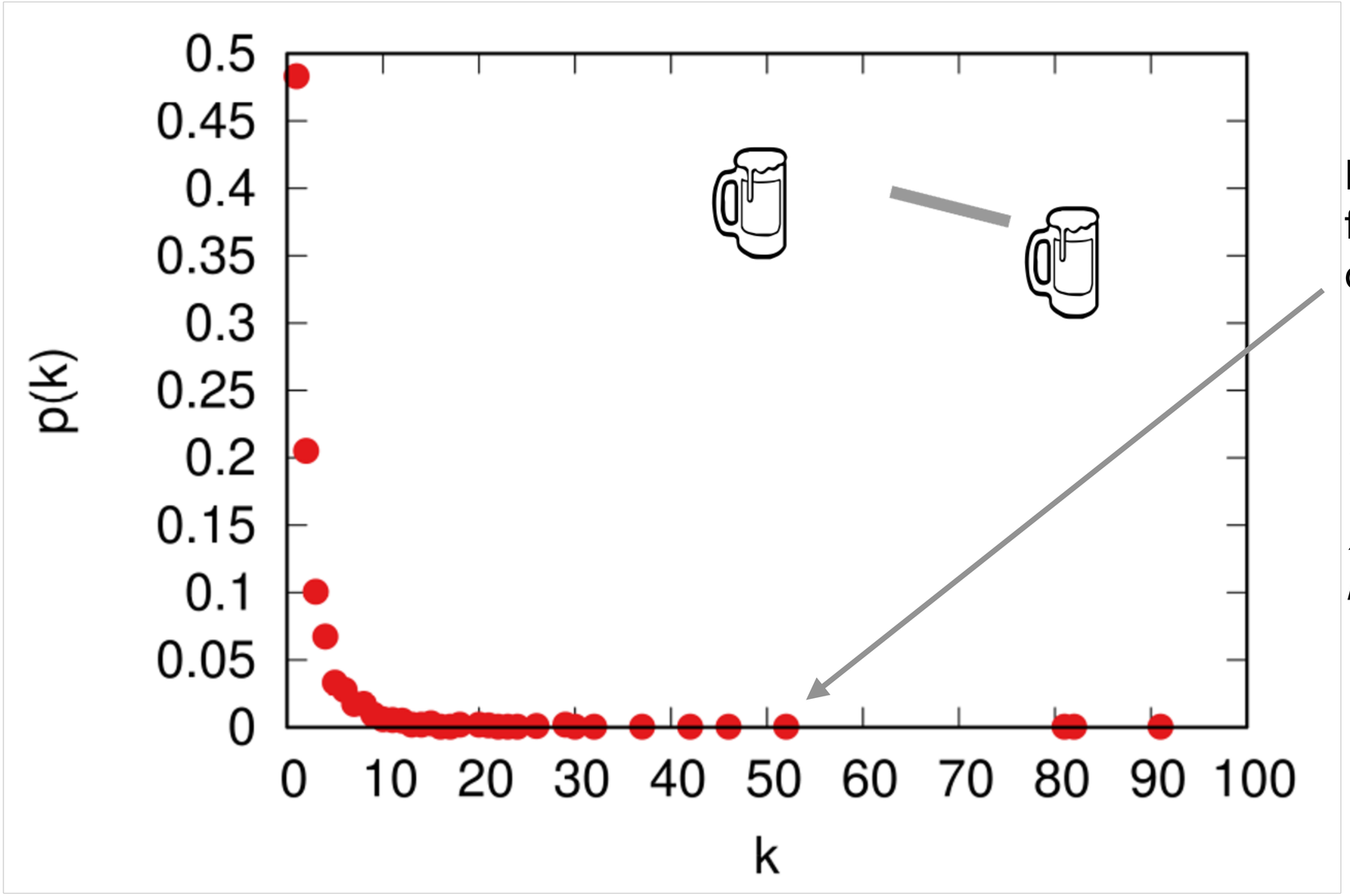
Better using scatter plots or curves

Real degree distribution



Protein-protein interaction for the Saccharomyces Cerevisiae, the beer bug.

Real degree distribution

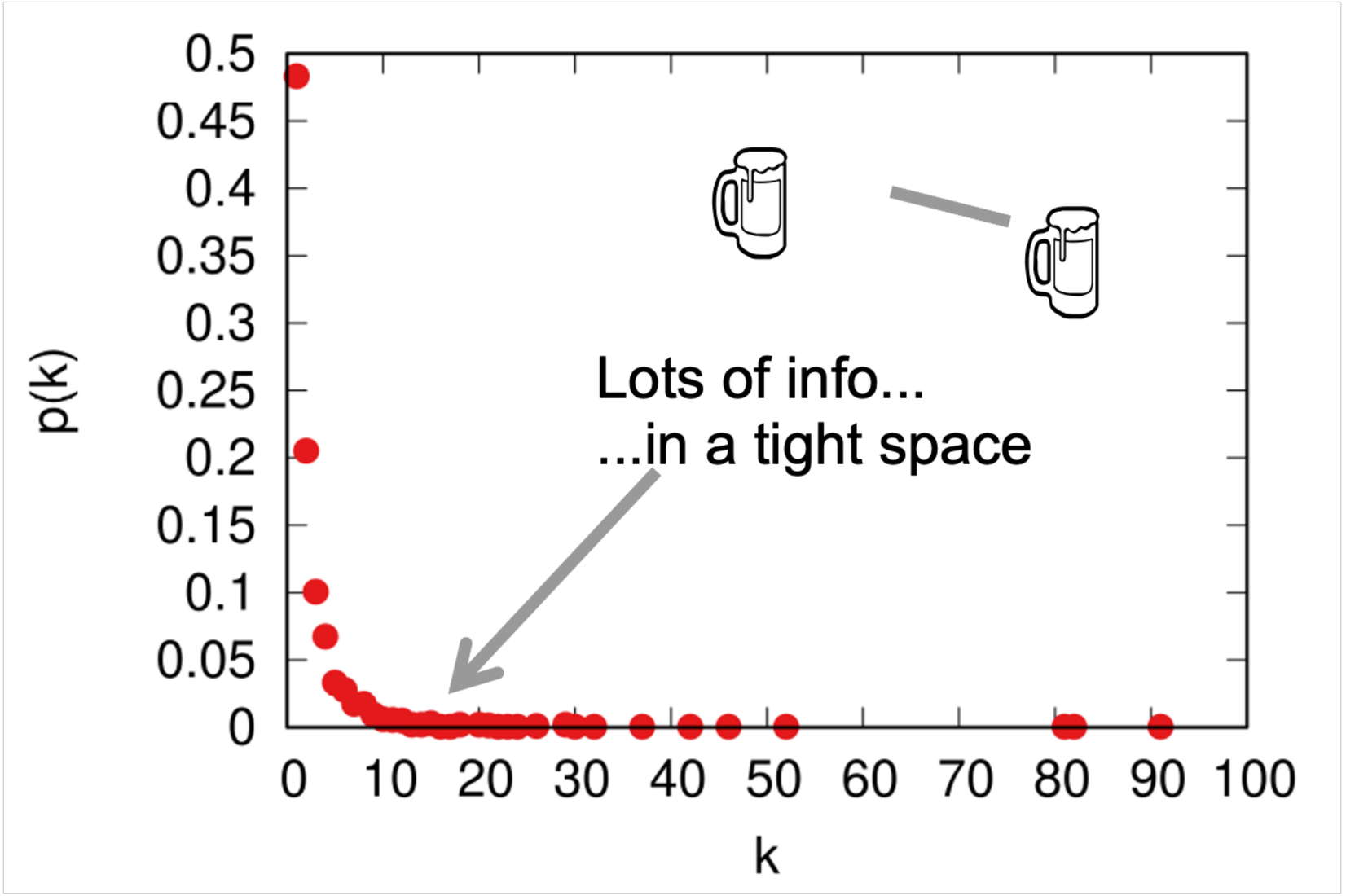


Probability of finding a node of degree **k**

$$\sum_{k=1}^{\infty} p_k = 1$$

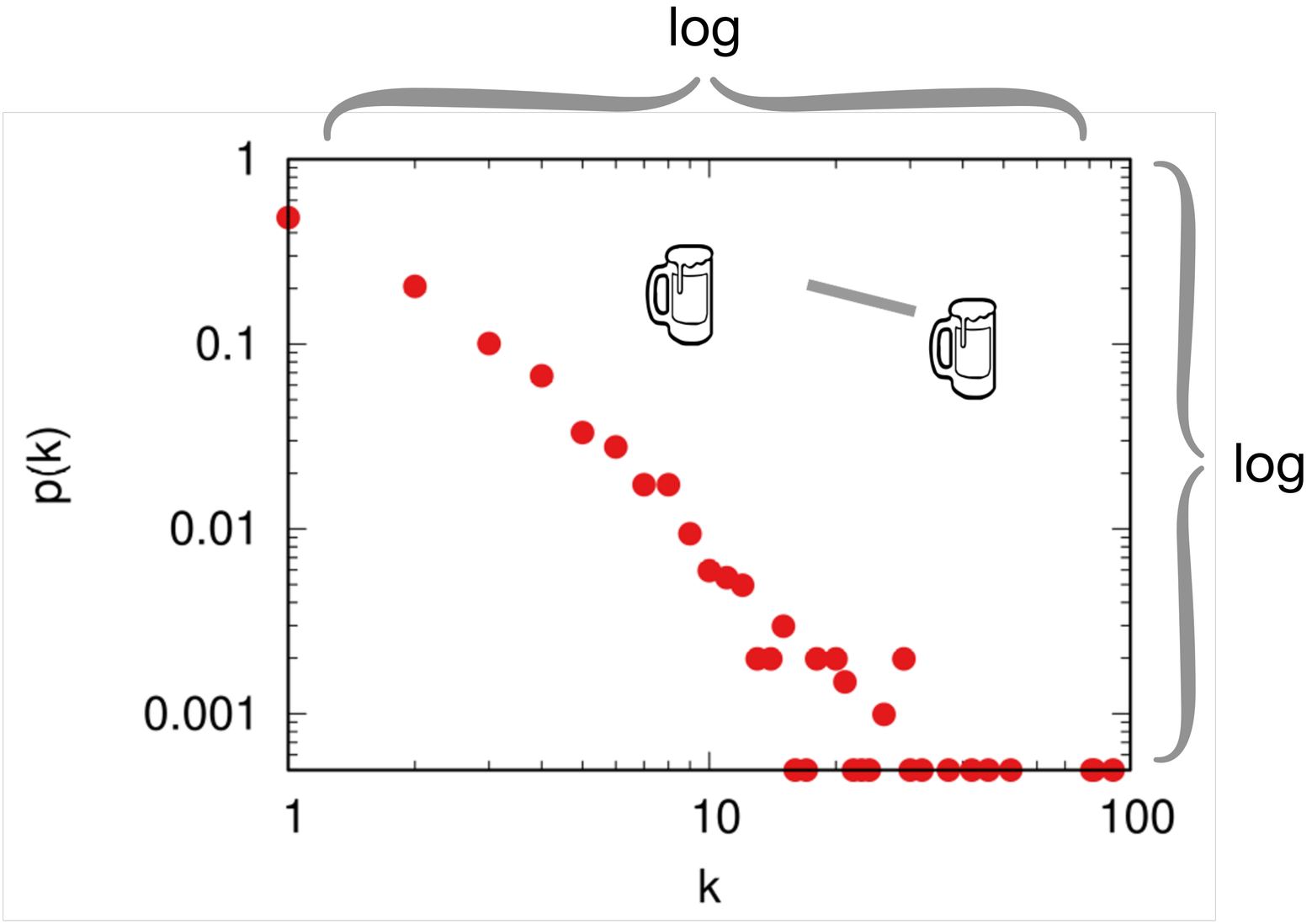
Protein-protein interaction for the Saccharomyces Cerevisiae, the beer bug.

Real degree distribution



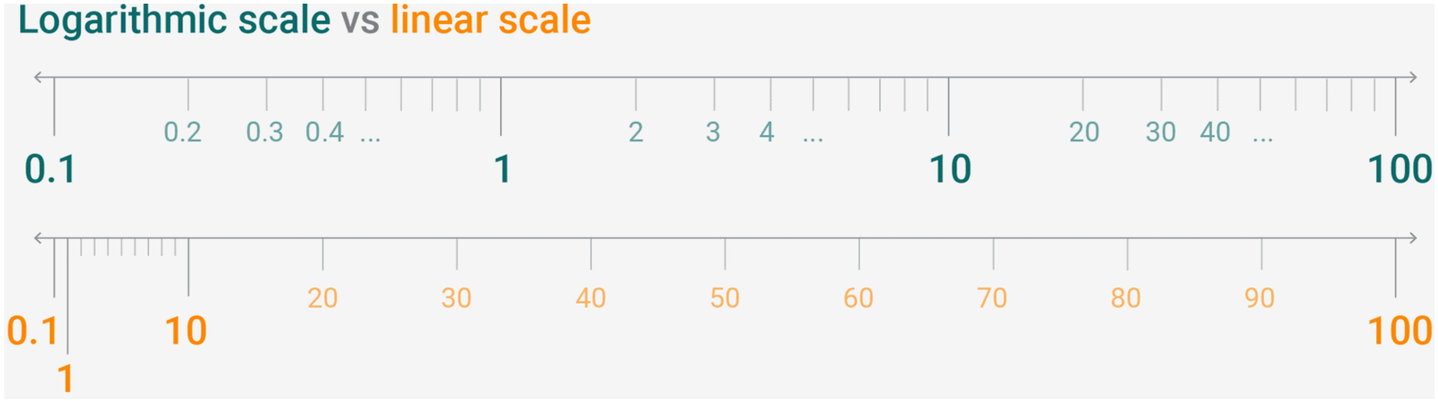
Protein-protein interaction for the Saccharomyces Cerevisiae, the beer bug.

Real degree distribution



Side note: Logarithmic scale

Question: how to plot a probability distribution if the variable spans a large range of values, from small to (very) large?

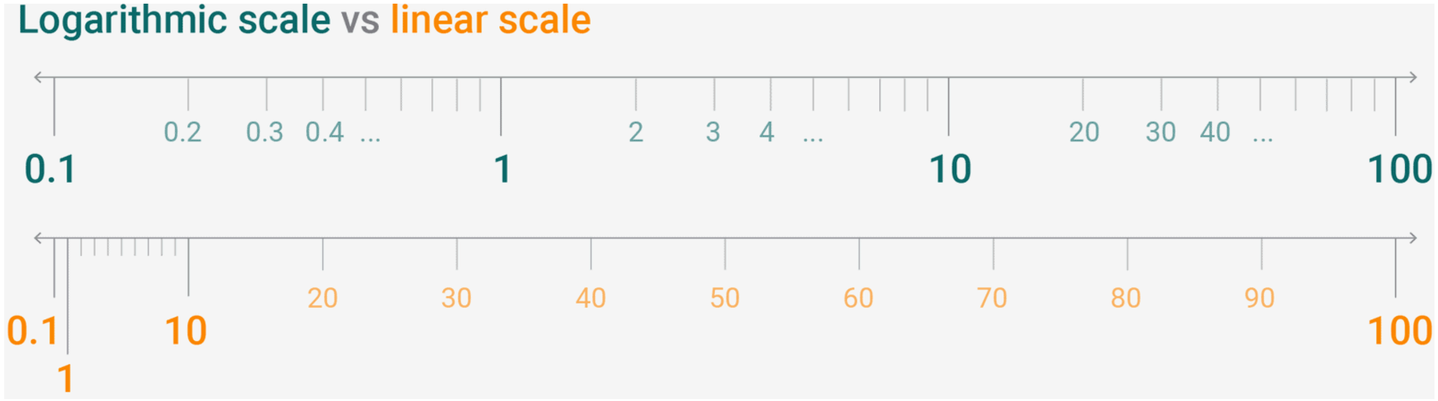


Distance between 0.1 and 1 is as big as the distance between 1 and 10 and 100,000 and 1,000,000.

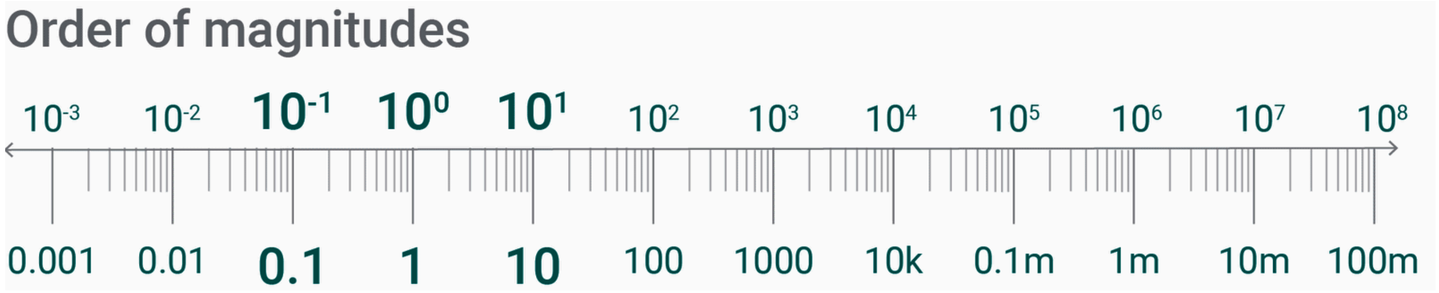
From: https://blog.datawrapper.de/weeklychart-logscale2/

Side note: Logarithmic scale

Question: how to plot a probability distribution if the variable spans a large range of values, from small to (very) large?



Distance between 0.1 and 1 is as big as the distance between 1 and 10 and 100,000 and 1,000,000.

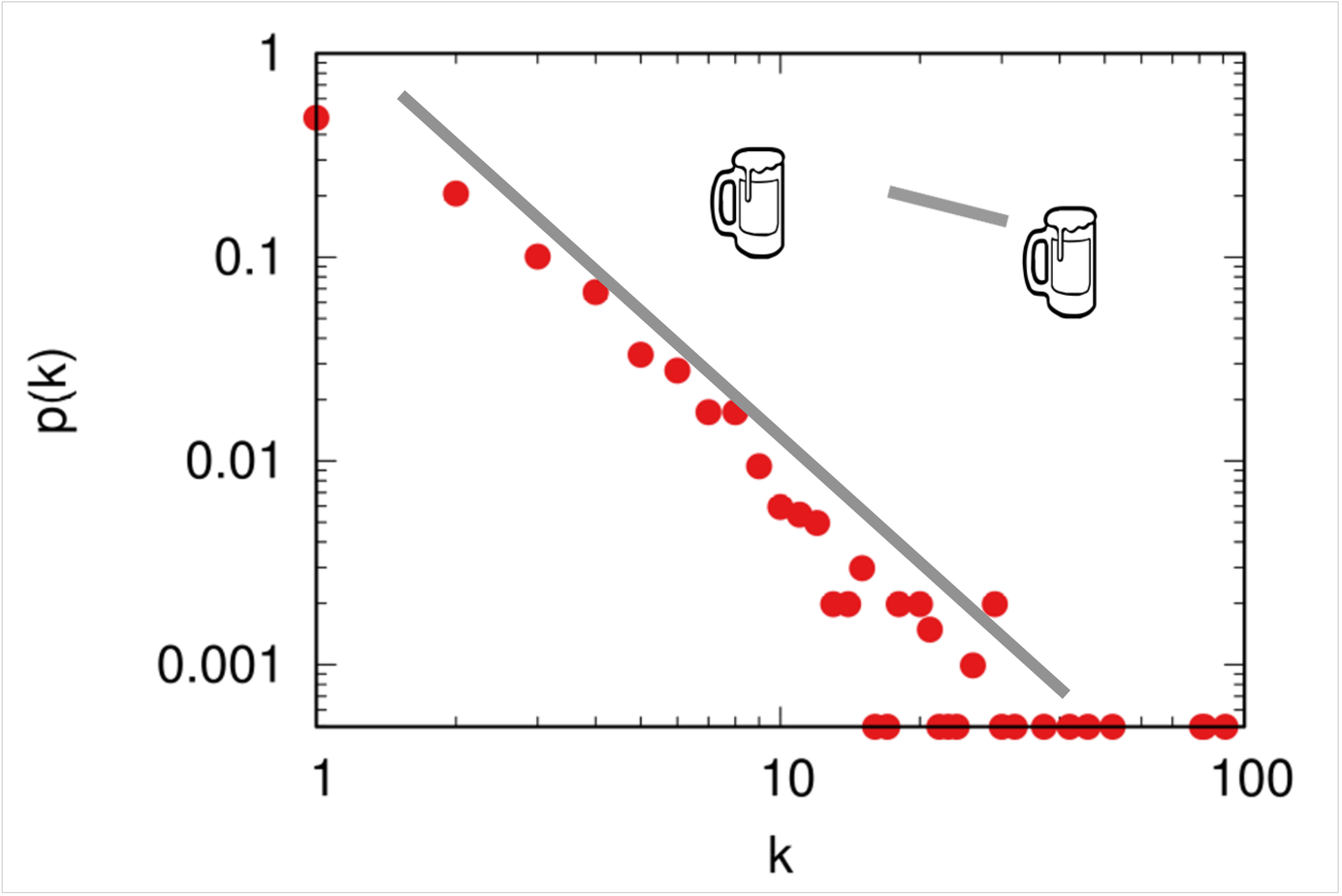


Common to report the logarithm (exponent) on the x and y-axes

From: https://blog.datawrapper.de/weeklychart-logscale2/

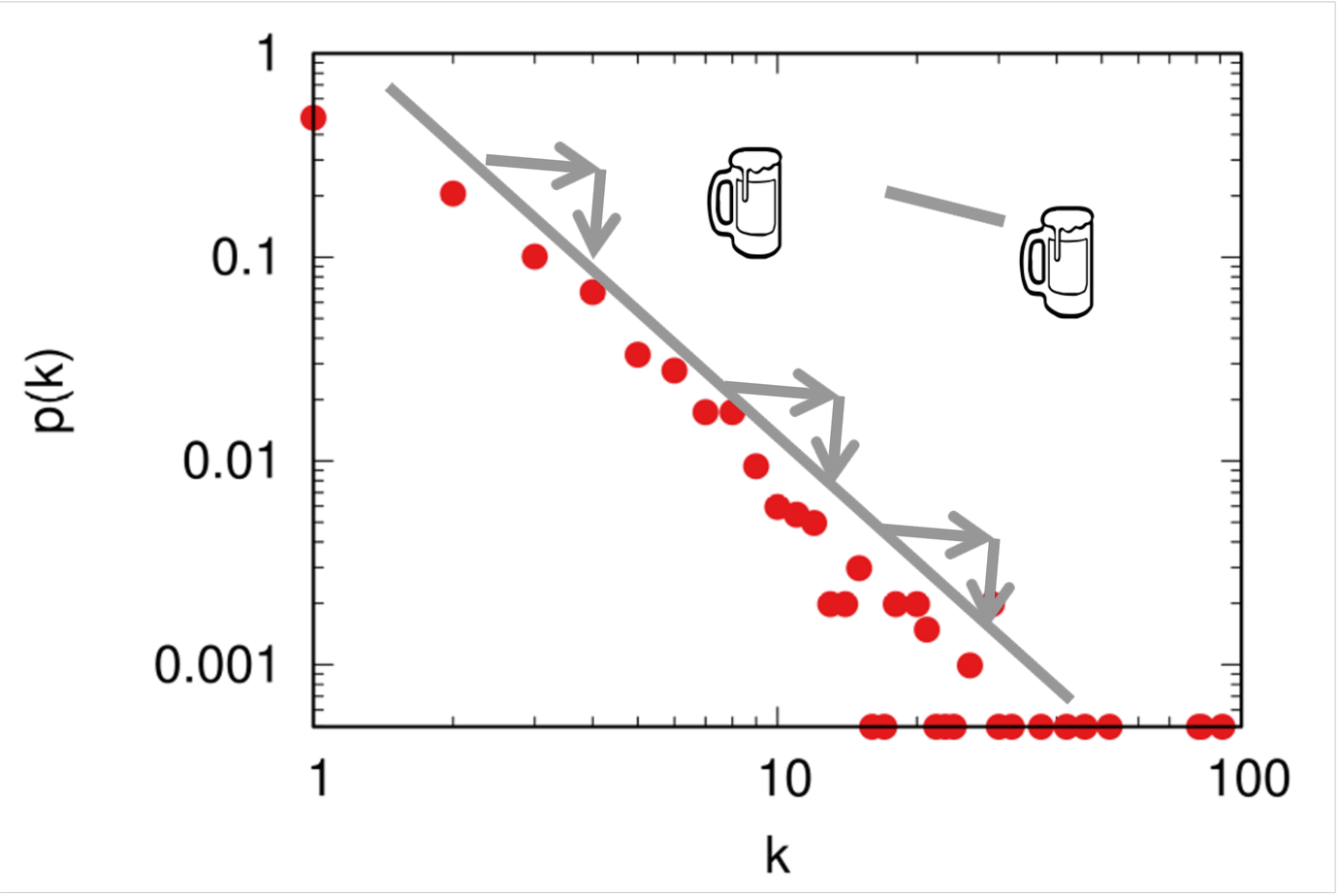
Real degree distribution

What approximating the points by a straight line here mean?



Real degree distribution

What approximating the points by a straight line here mean?



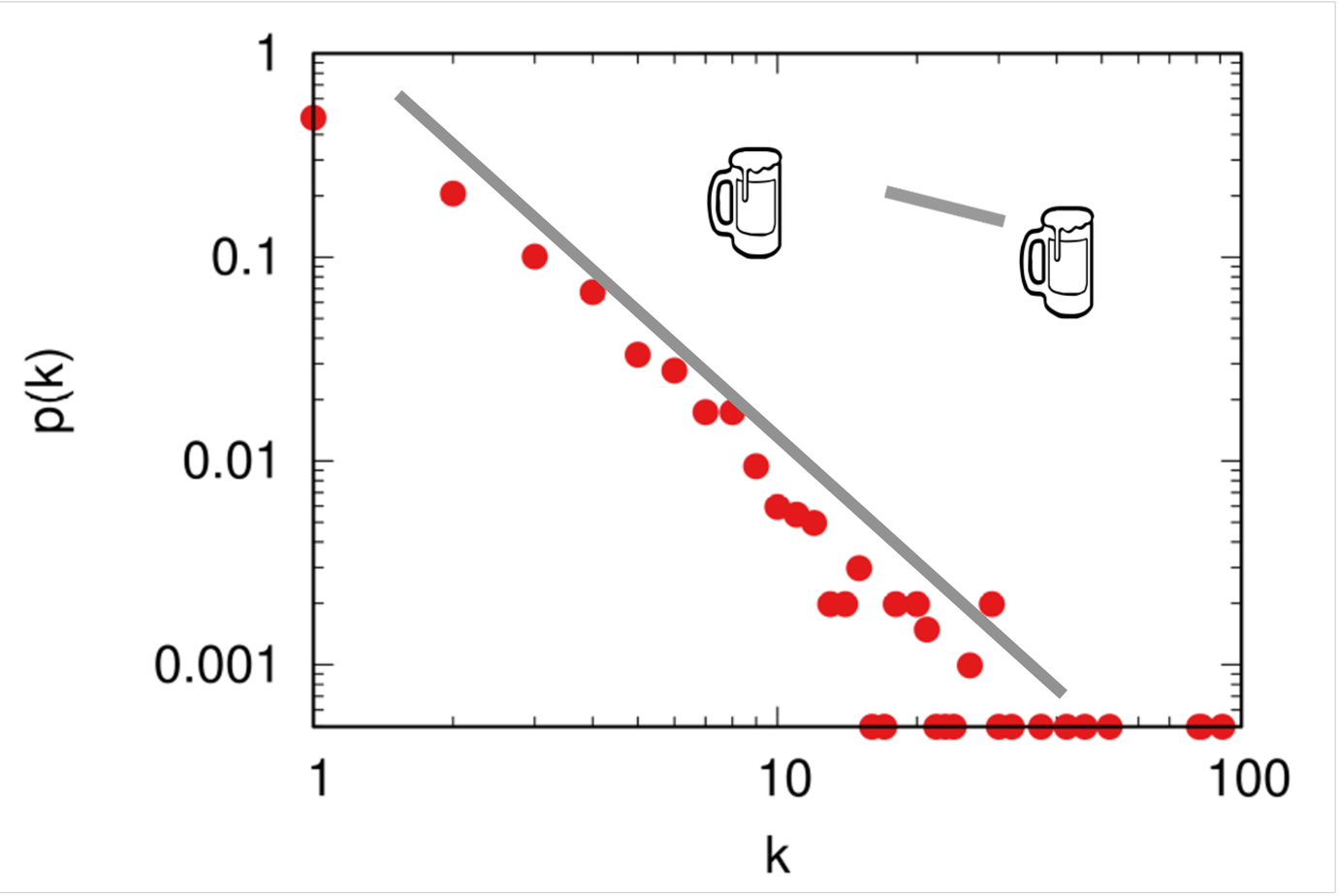
Relative change in one quantity

Proportional relative change in the other

Independent of their initial size

Real degree distribution

What approximating the points by a straight line here mean?



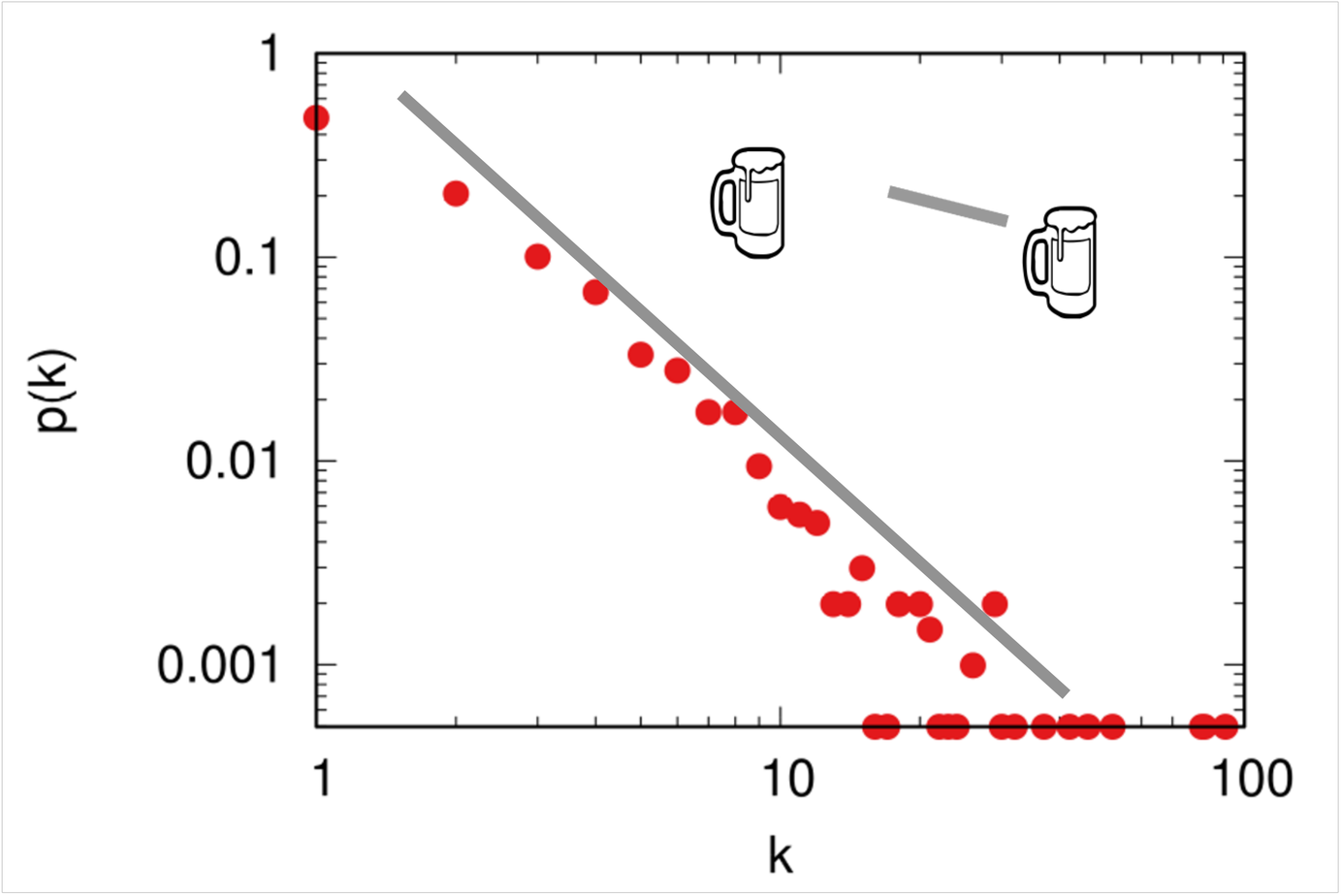
Linear relationship

A straight line in log-log —> the degree distribution is approximated by

$$p(k) \sim k^{-\gamma}$$

Real degree distribution

What approximating the points by a straight line here mean?



Linear relationship

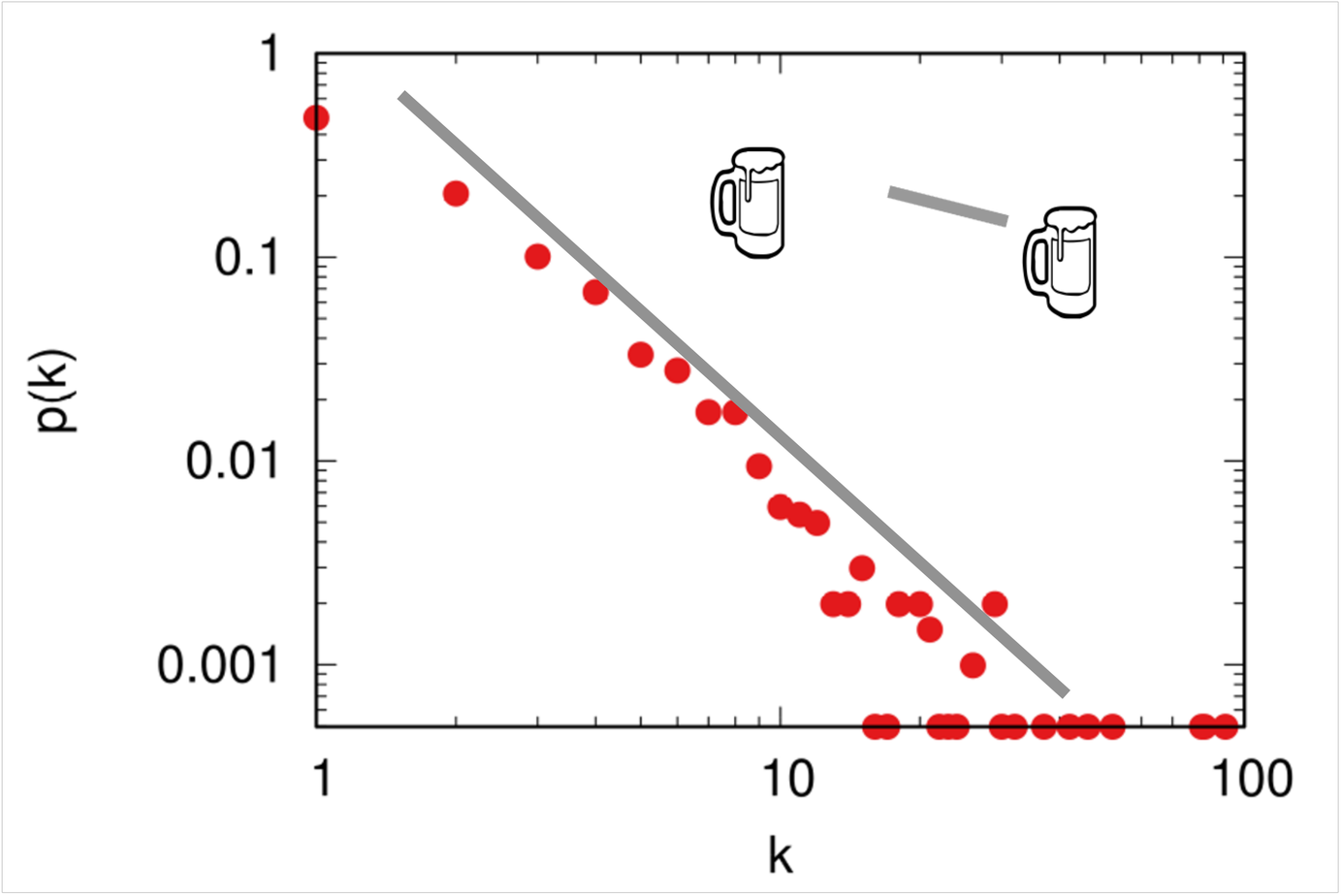
A straight line in log-log —> the degree distribution is approximated by

$$p(k) \sim k^{-\gamma}$$

The probability of a node having degree ***k*** is ***k*$^{\gamma}$**,** and the exponent γ is the slope of the line (degree exponent)

Real degree distribution

What approximating the points by a straight line here mean?



Linear relationship

A straight line in log-log —> the degree distribution is approximated by

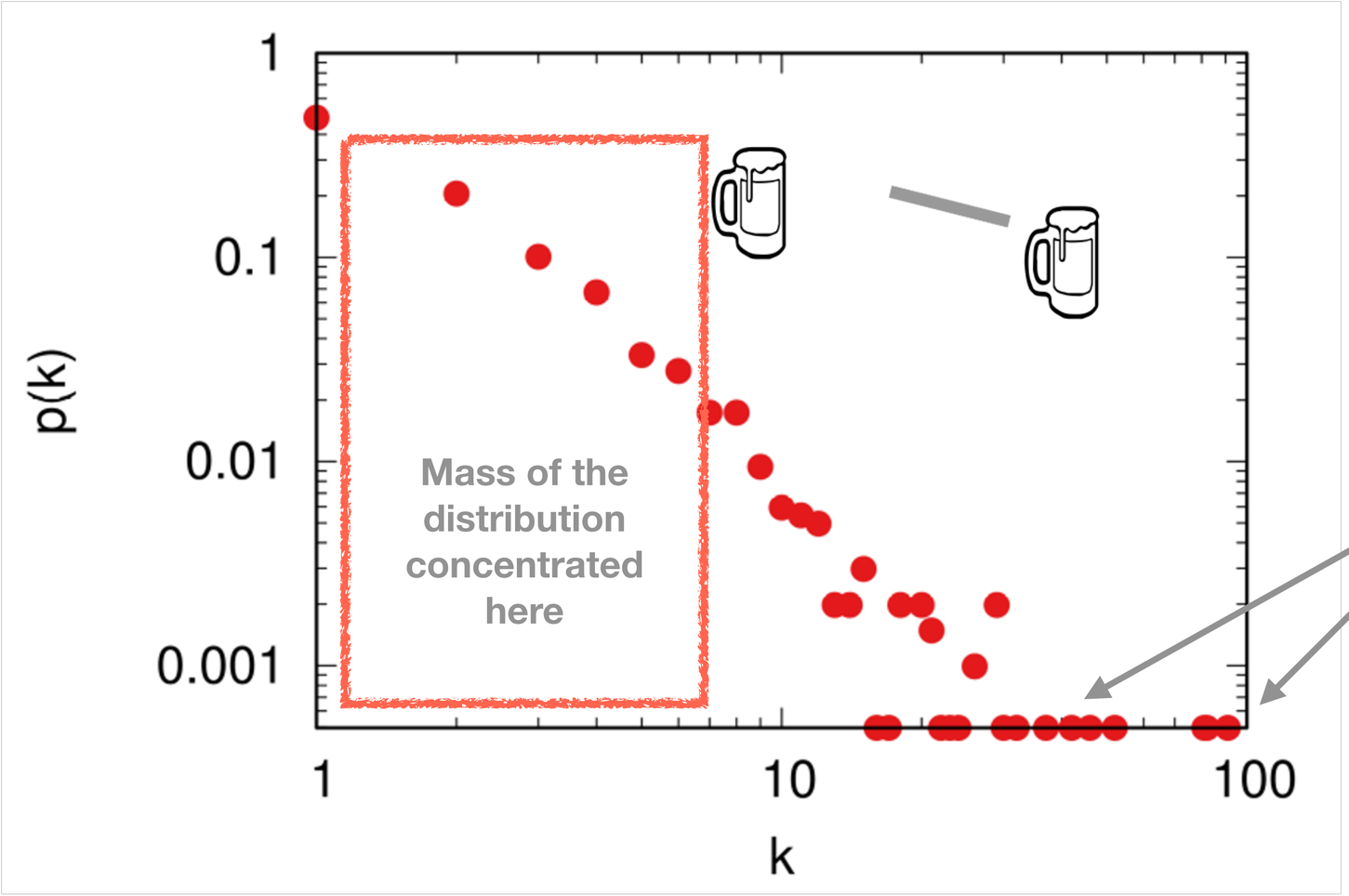
$$p(k) \sim k^{-\gamma}$$

The probability of a node having degree ***k*** is ***k*$^{\gamma}$**, and the exponent γ is the slope of the line (degree exponent)

This is called a power law, surprise!

Not covered here, it is a super cool property of many networks in nature. Check Ch.4 of the book if you are curious.

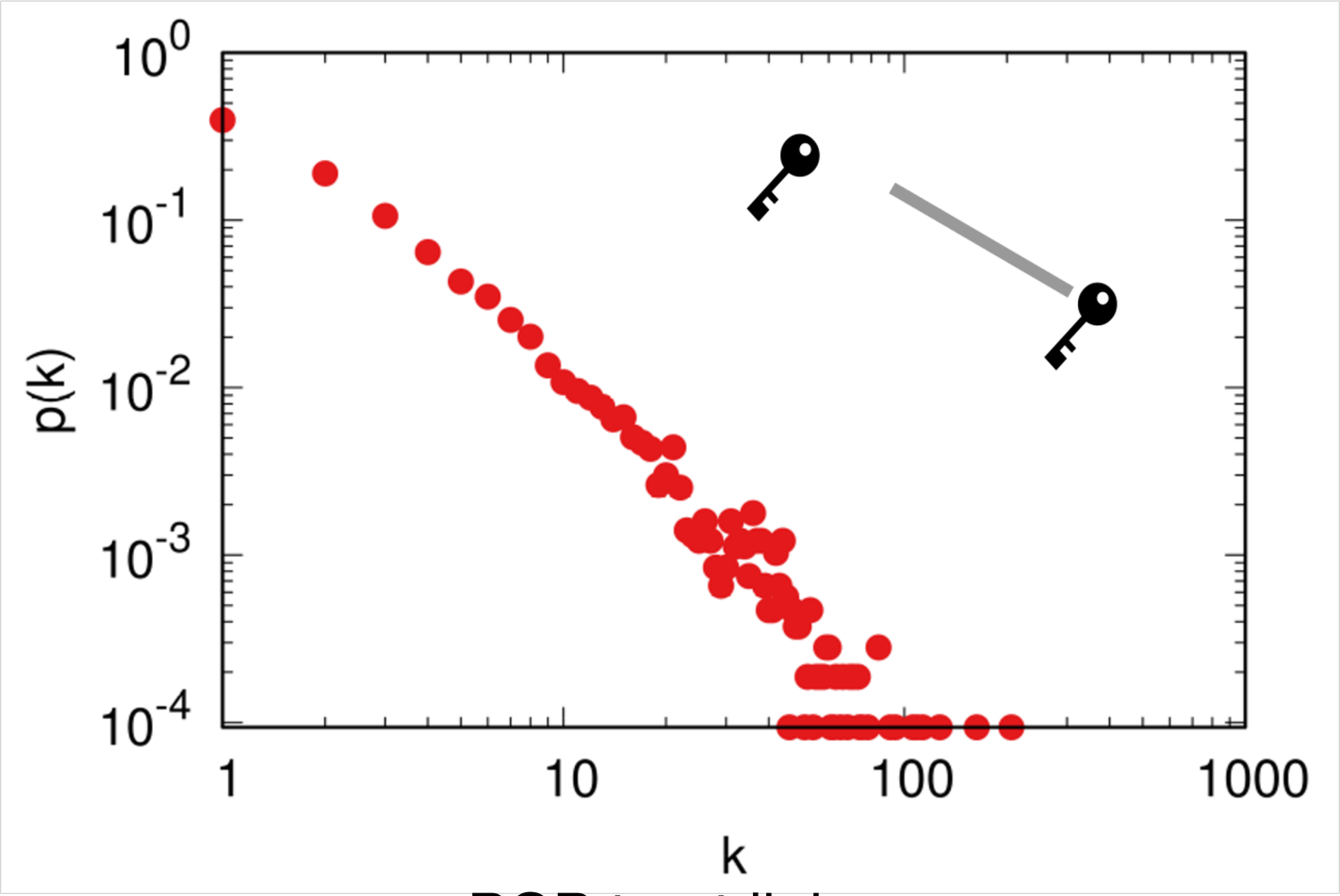
Real degree distribution



Heavy-tail (fat-tail) distributions: the variable goes from small to large values

Real degree distribution

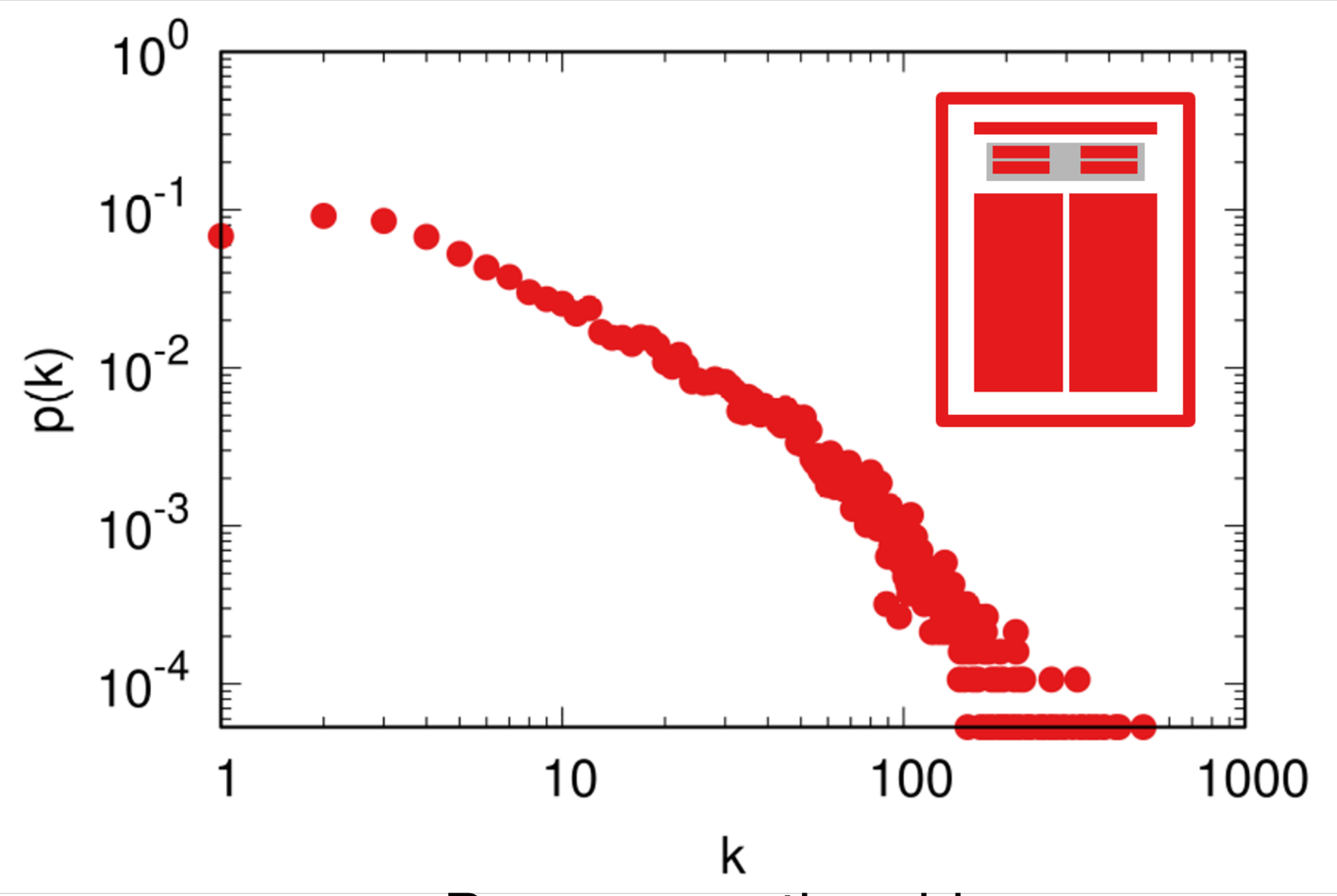
The vast majority of networks have broad degree distributions, spanning multiple orders of magnitude.



PGP trust links

Real degree distribution

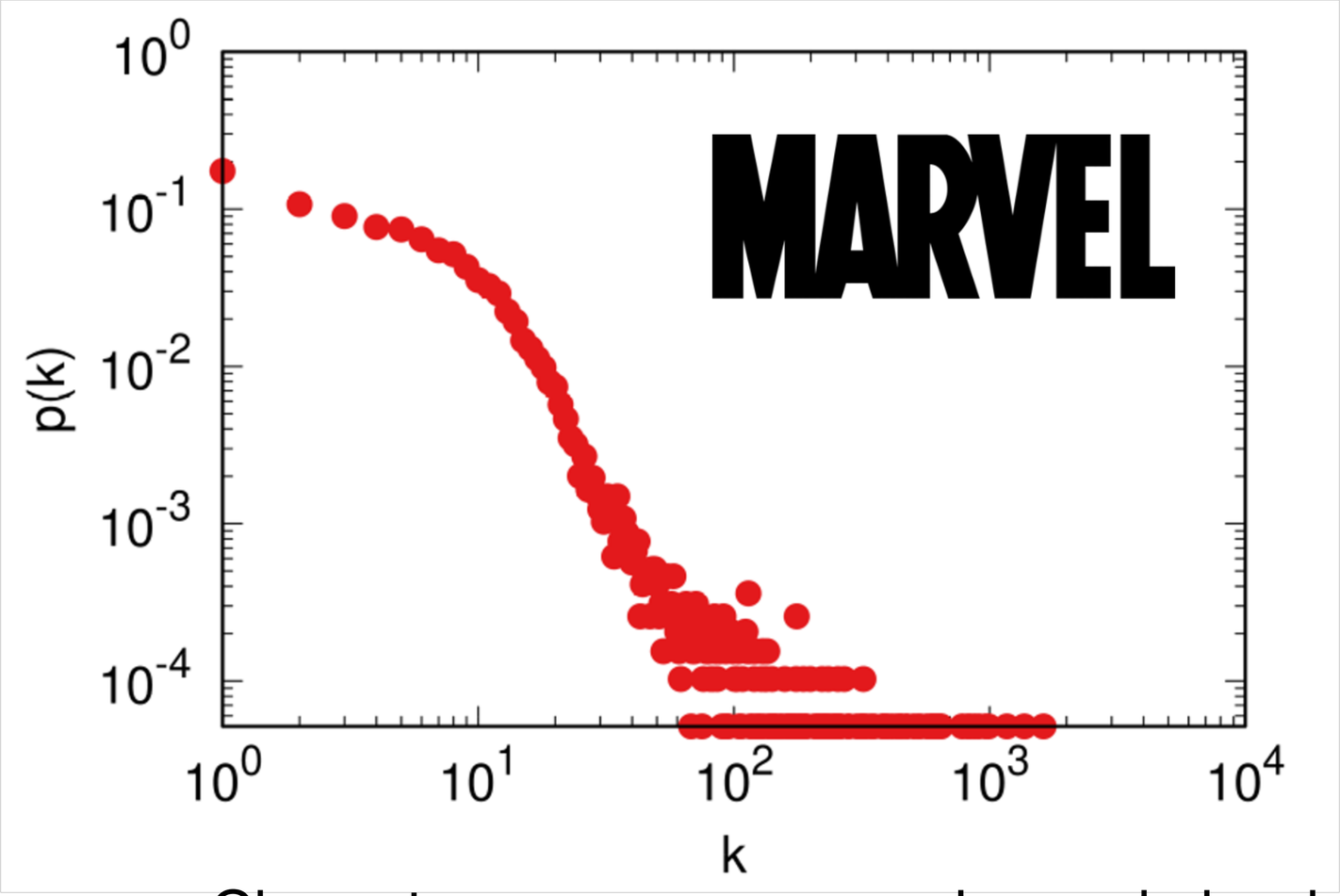
Most nodes have below-average degrees and we have hubs many standard deviations above the average.



Paper co-authorship

Real degree distribution

Not all networks will have a clear relationship of ***p(k)*** and ***k*** approximated by a straight line.



Character co-appearances in comic books

Degree distribution: formalising

The degree distribution, ***pk***, provides the probability that a randomly selected node in the network has degree ***k***. Since ***pk*** is a probability, it must be normalized, i.e.

$$\sum_{k=1}^{\infty} p_k = 1$$

Degree distribution: formalising

The degree distribution, ***pk***, provides the probability that a randomly selected node in the network has degree ***k***. Since ***pk*** is a probability, it must be normalized, i.e.

$$\sum_{k=1}^{\infty} p_k = 1$$

For a network with **N** nodes, the degree distribution is the normalised histogram given by:

$$p_k = \frac{N_k}{N}$$

where N_k is the number of degree-k nodes.

Degree distribution: formalising

The degree distribution, ***pk***, provides the probability that a randomly selected node in the network has degree ***k***. Since ***pk*** is a probability, it must be normalized, i.e.

$$\sum_{k=1}^{\infty} p_k = 1$$

For a network with **N** nodes, the degree distribution is the normalised histogram given by:

$$p_k = \frac{N_k}{N}$$

where N_k is the number of degree-k nodes.

Hence, the number of degree-k nodes can be obtained from the degree distribution as $N_k = Np_k$.

Degree distribution: formalising

The degree distribution, ***pk***, provides the probability that a randomly selected node in the network has degree ***k***. Since ***pk*** is a probability, it must be normalized, i.e.

$$\sum_{k=1}^{\infty} p_k = 1$$

For a network with **N** nodes, the degree distribution is the normalised histogram given by:

$$p_k = \frac{N_k}{N}$$

where N_k is the number of degree-k nodes.

Hence, the number of degree-k nodes can be obtained from the degree distribution as $N_k = Np_k$.

We can write the average degree of a network as:

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k$$

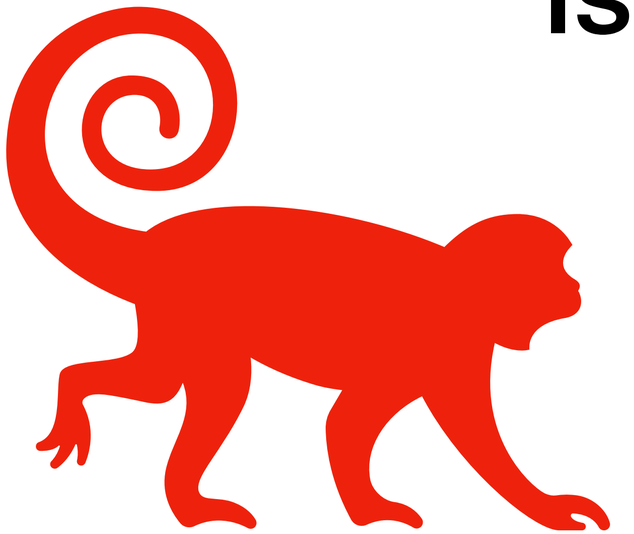
Random network model

Why do we model?

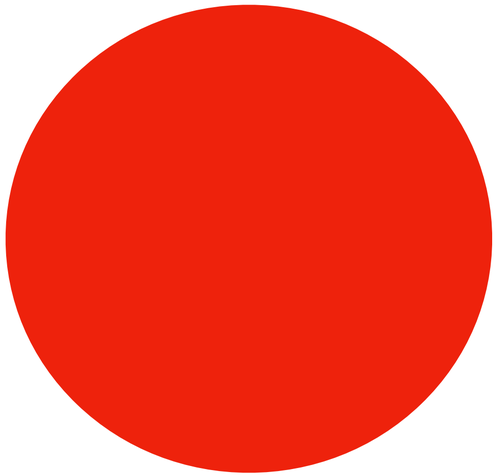
"All models are wrong but some are useful" (George Box, 1978, p. 202-3)

Why do we model?

Is this reasonable?



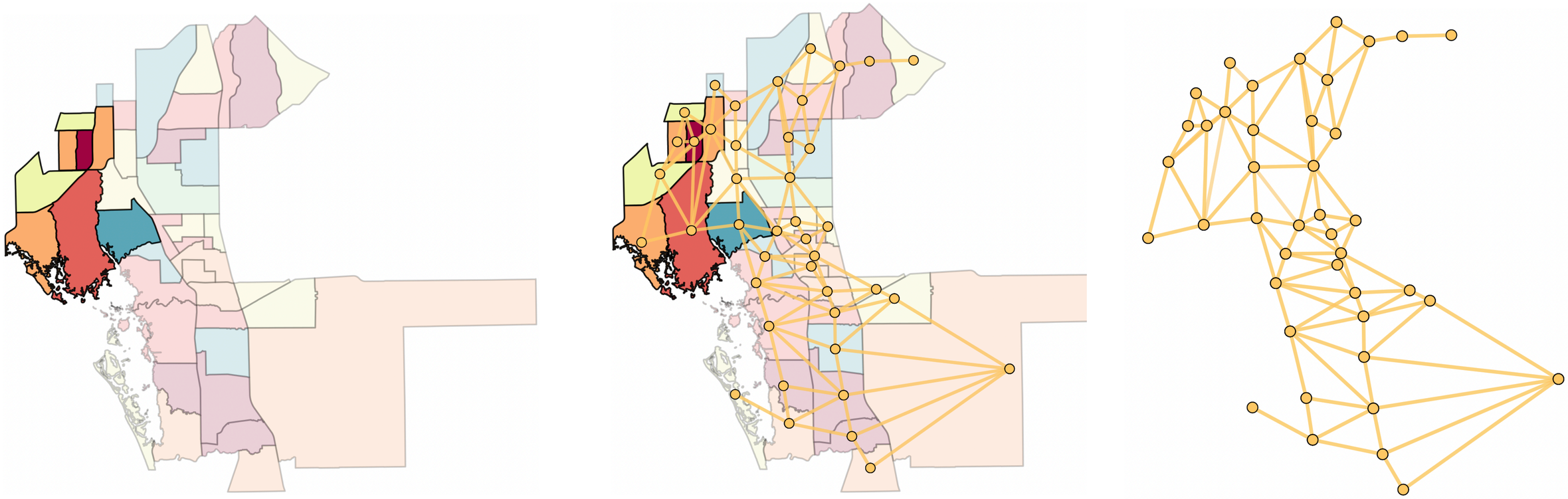
Modelling



YES and NO

Why do we model?

Coarse graining a problem into meaningfully workable and simplified parts.
A network is a model of a real-world system, it is not the system.

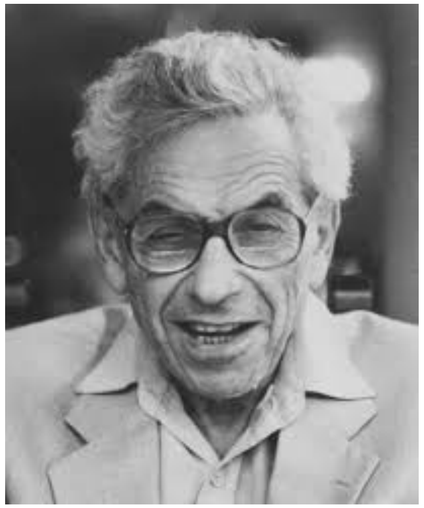


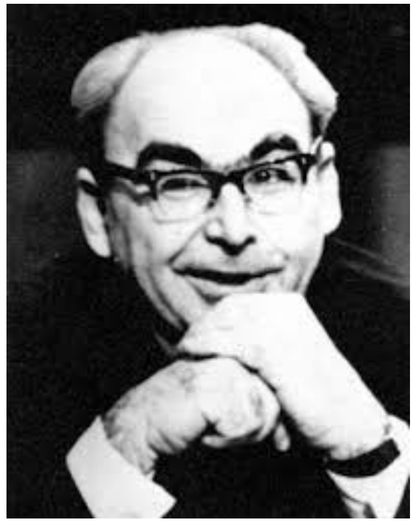
This model is as helpful as the questions we ask about the system.

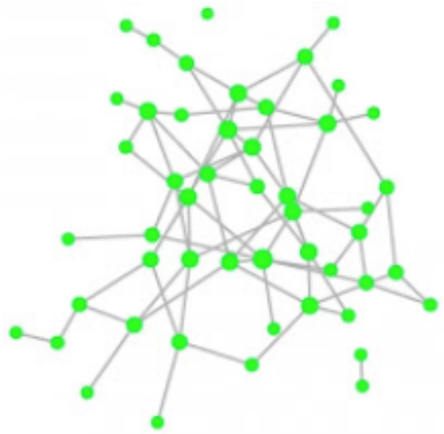
Bassolas, A., Sousa, S. and Nicosia, V., 2021. Journal of the Royal Society Interface, 18(174), p.20200961

Network models

Models before massive data was available

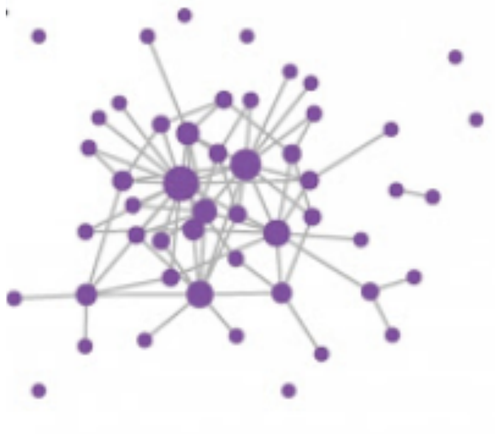






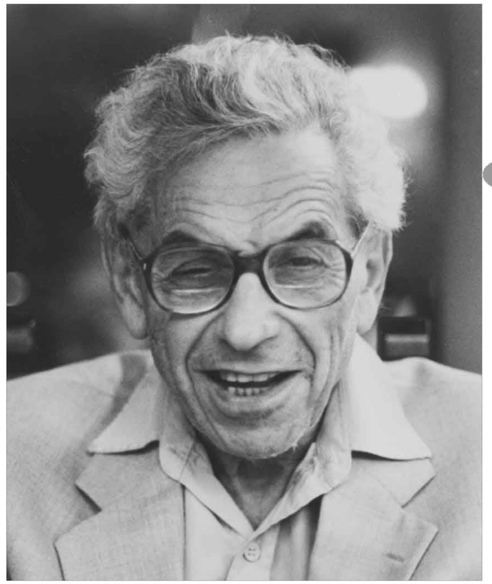
Models after massive data was available

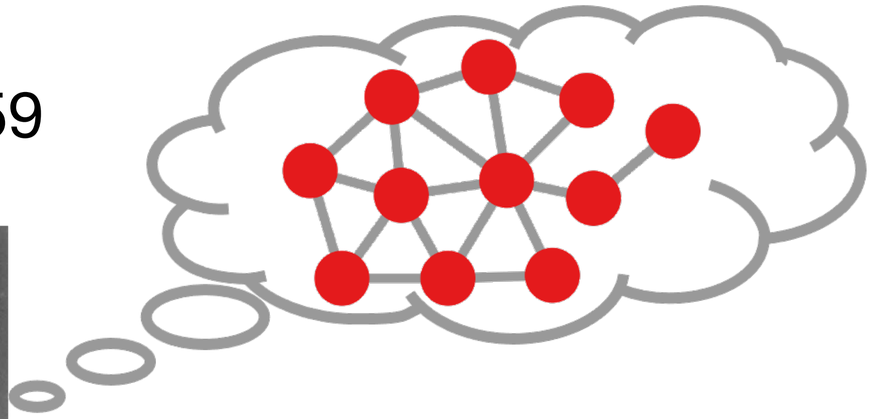
ID	BILL DATE	VENDOR	BAN	MATCH TO	MATCH STATUS	AUDIT
1	11/8/2008 0:00:00	ATT-NRC	3107020003552	REF ONLY	EXCEPT	
2	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	EX	
3	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	FC	
4	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	EX	
5	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	FC	
6	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	FC	
7	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	FC	
8	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	FC	
9	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	VA	
10	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
11	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
12	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
13	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
14	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
15	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
16	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
17	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
18	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	V	
19	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
20	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
21	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
22	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
23	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	
24	11/8/2008 0:00:00	ATT-NRC	3107020003552	INV TO MIROR	P	



Random Network Model

Paul Erdős in 1959



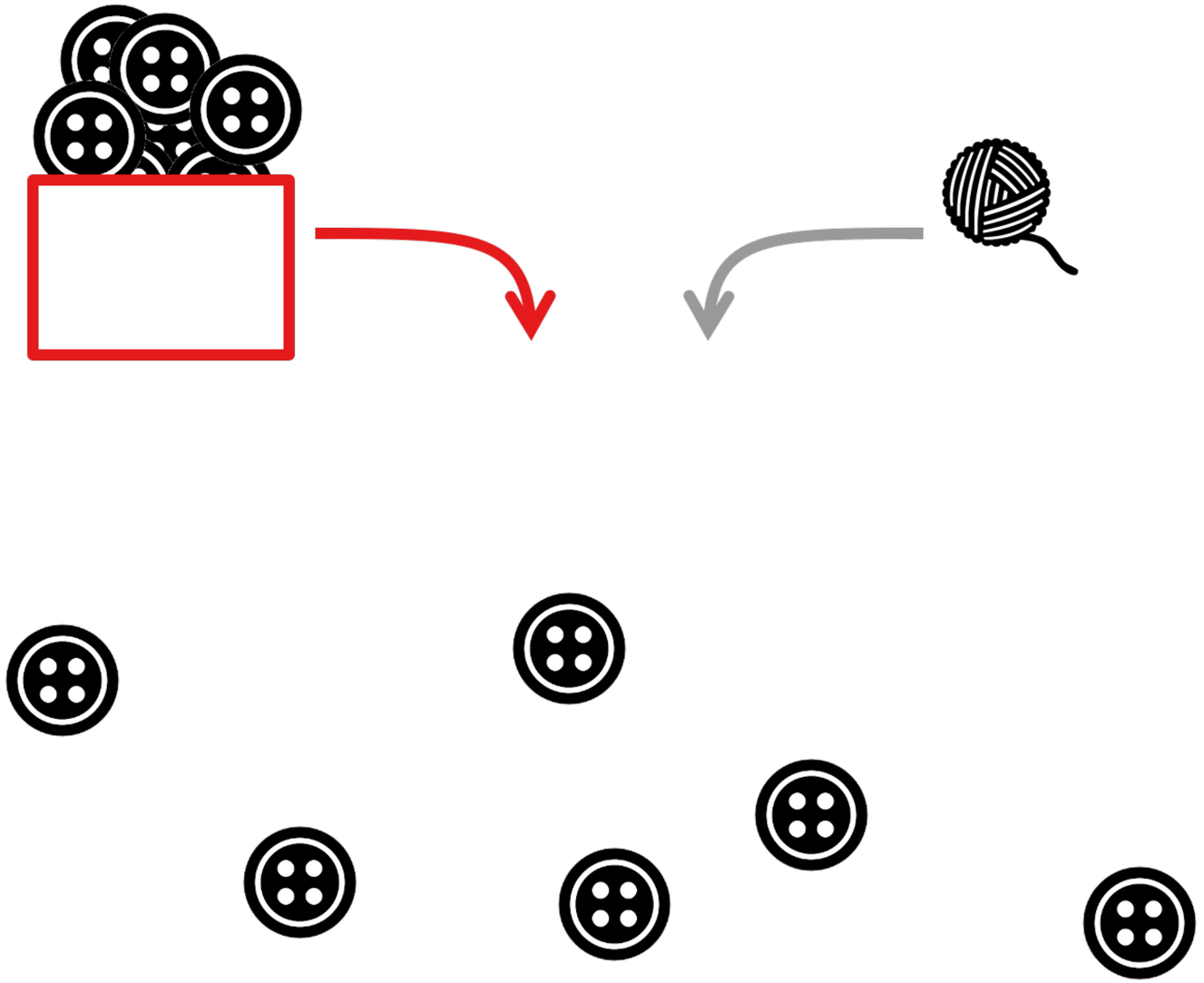


How do components emerge in networks?

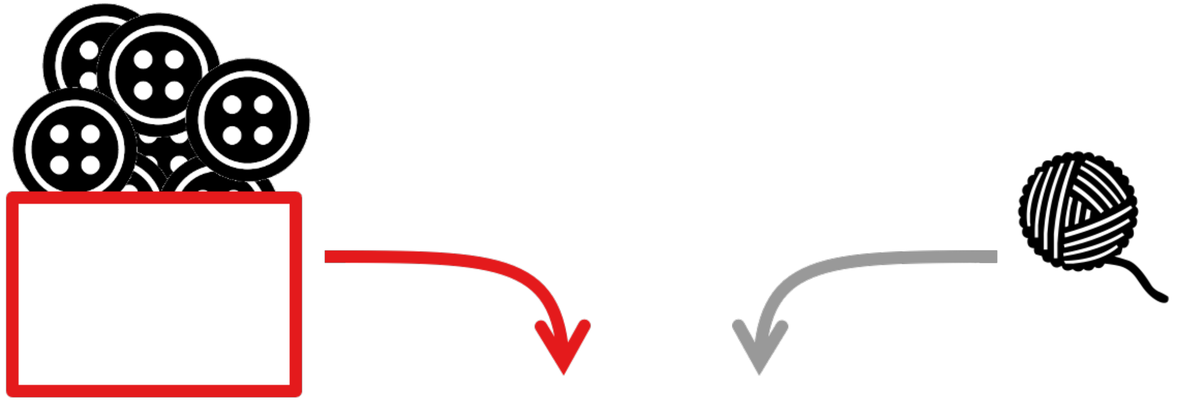
No computers
No data :/

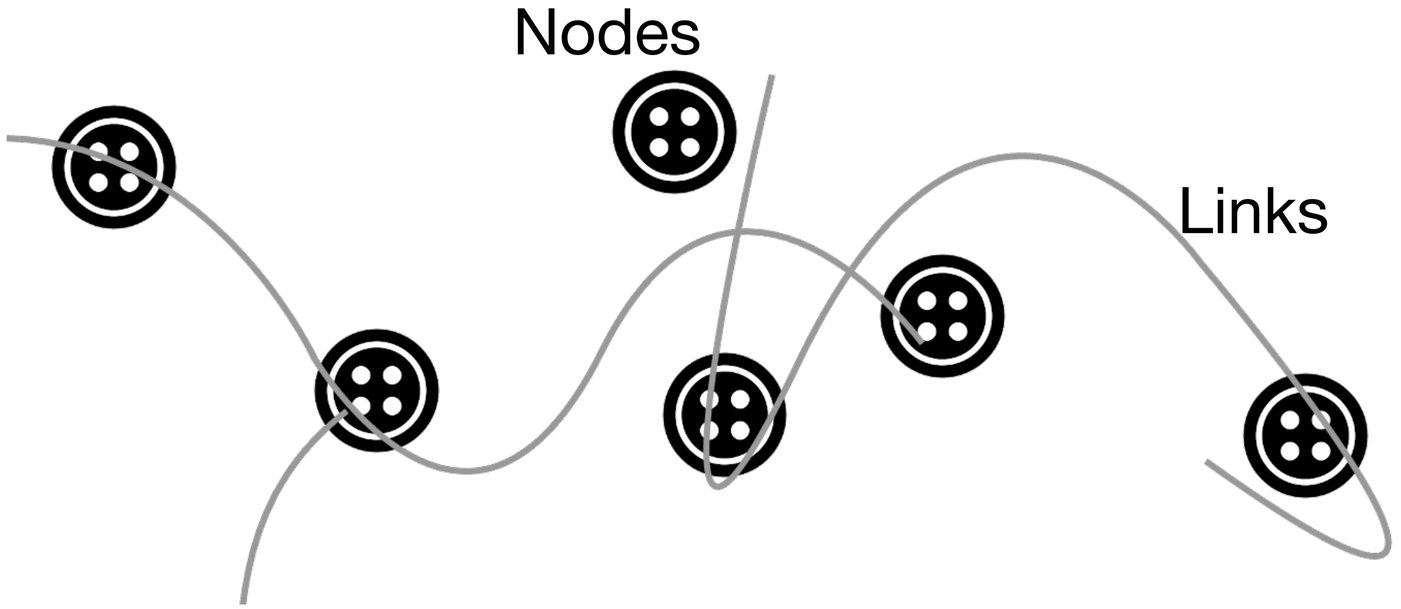
Let’s use pen and paper

Random Network Model



Random Network Model





Random Network Model

Random network models are truly random

You toss a die, the result is random

Random Network Model

Random network models are truly random

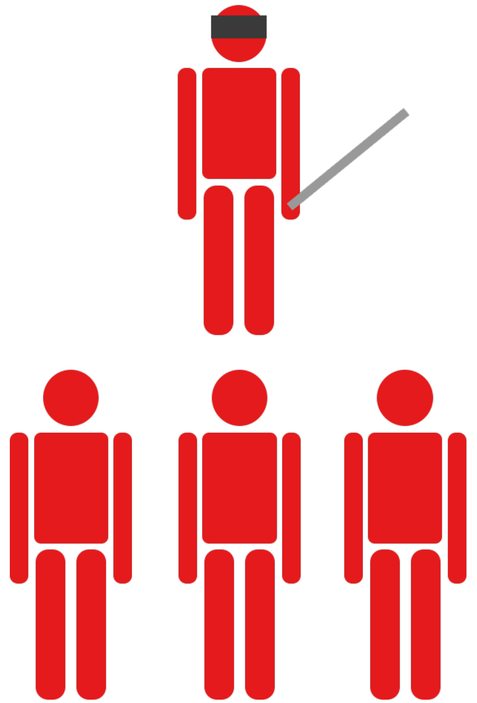
You toss a die, the result is random

In this model, random applies to the edges formation

Random Network Model

Random network models are truly random

You toss a die, the result is random



In this model, random applies to the edges formation

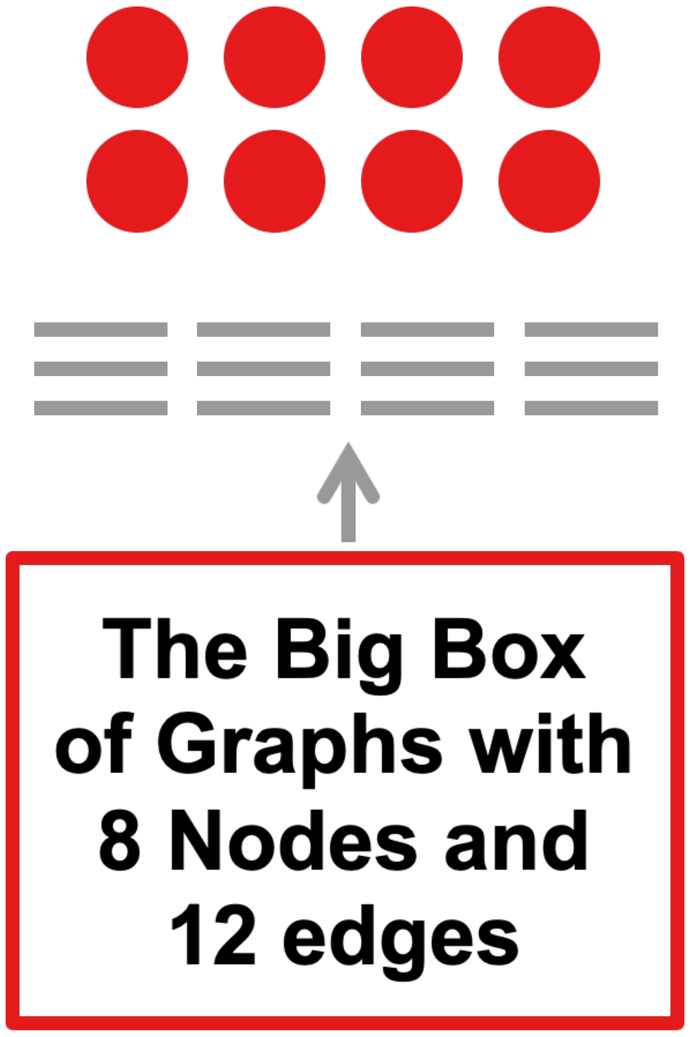
Random Network Model

G(n, m)

Fix # of nodes → n

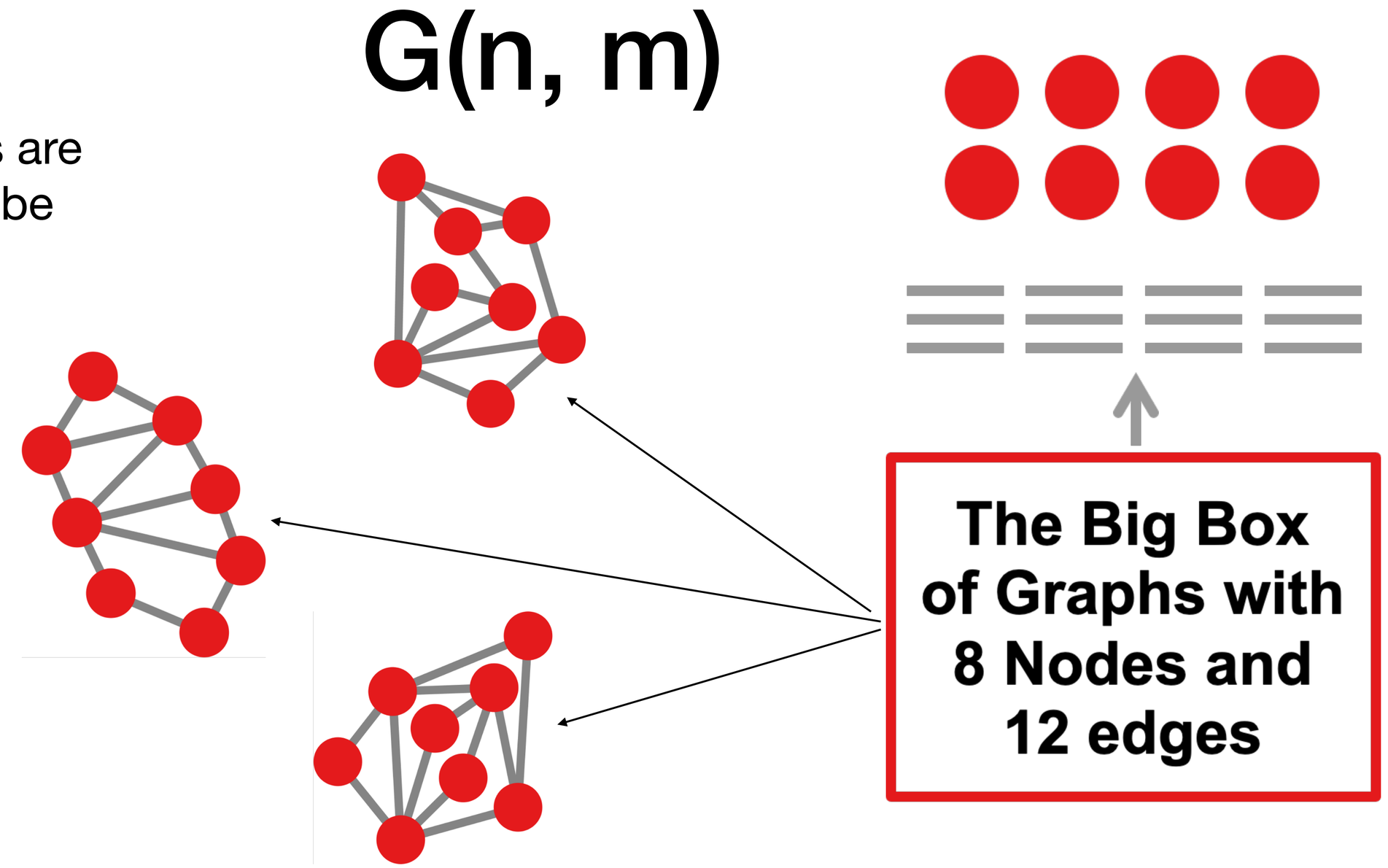
Fix # of edges → m

Pick a random graph with n nodes and m edges from the Big Box.



Random Network Model

All these graphs are equally likely to be selected.



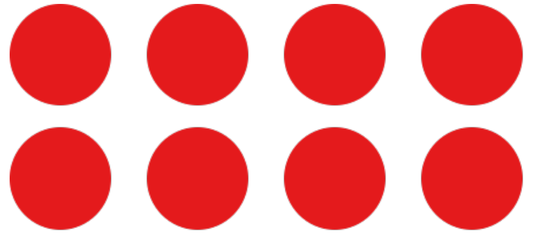
Random Network Model

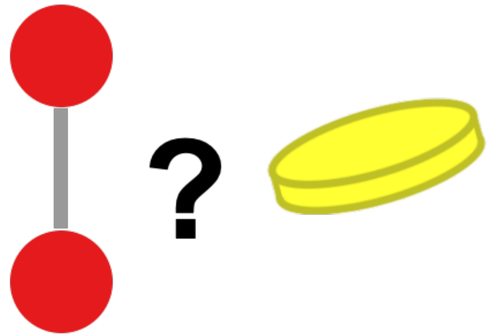
G(n, p)

Fix # of nodes → n

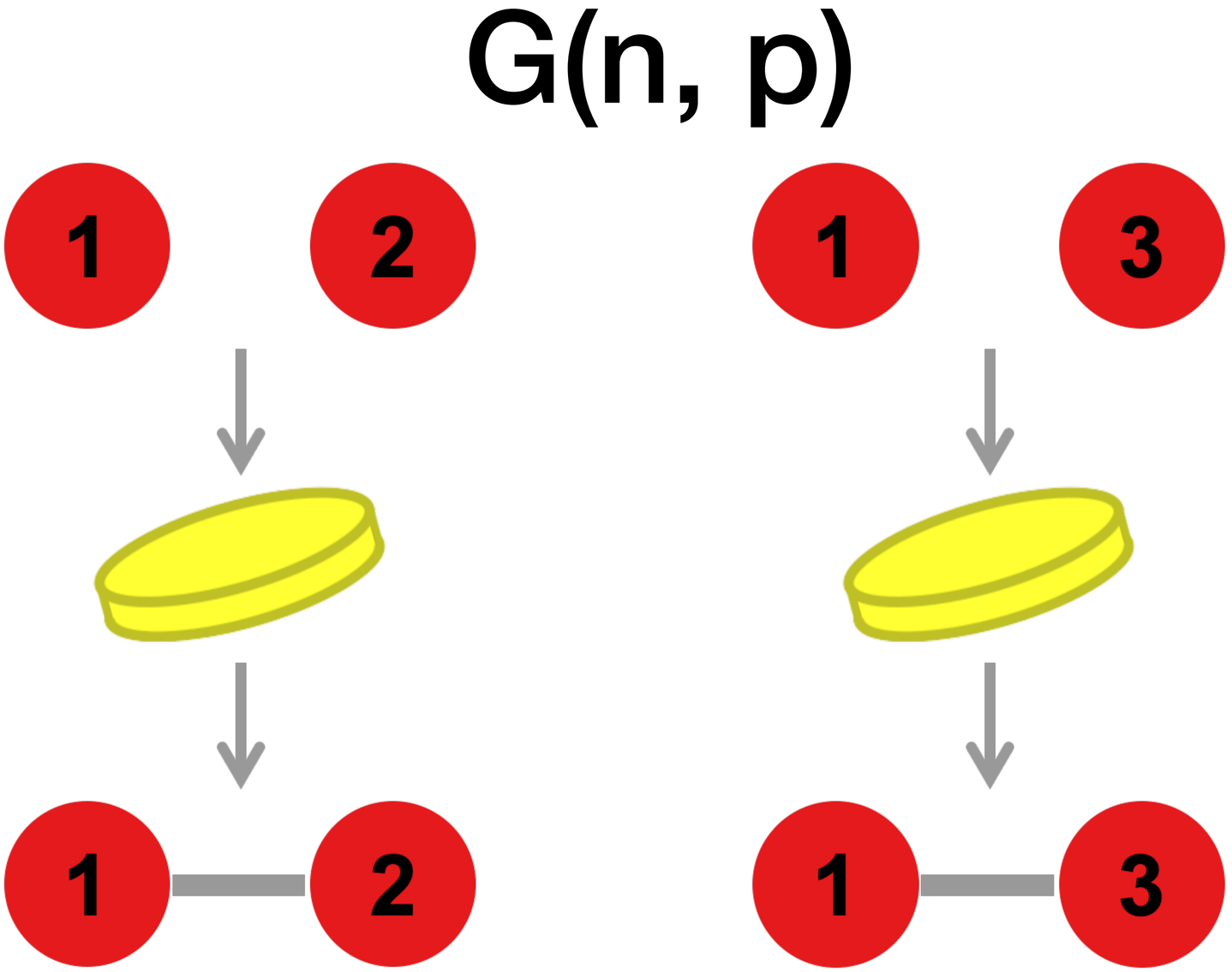
Fix connection probability → p

Check every possible
pair of nodes (construction
through iteration).

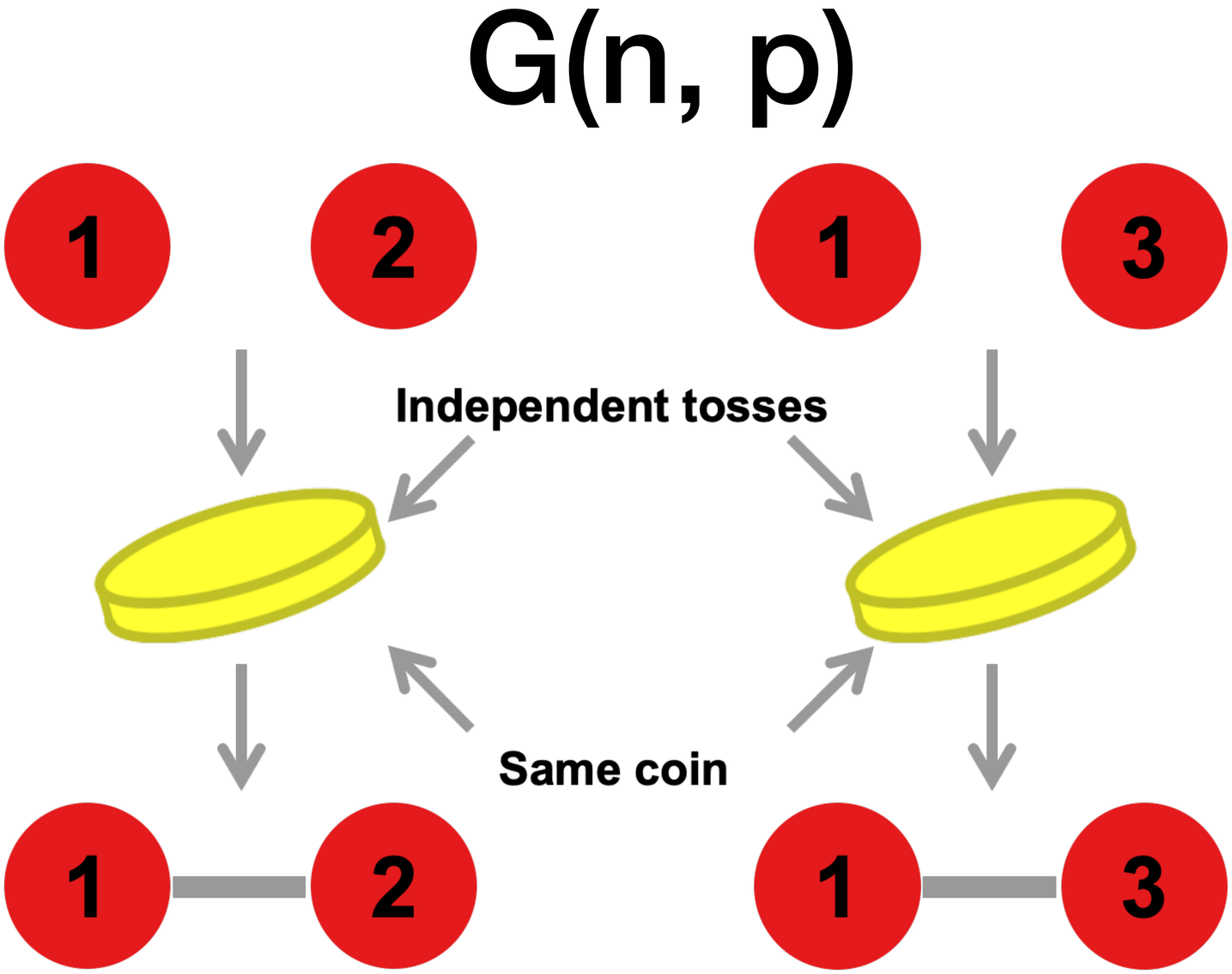




Random Network Model



Random Network Model



Random Network Model

Models p-m Equivalence

Referred as ER model (Erdős and Rényi)

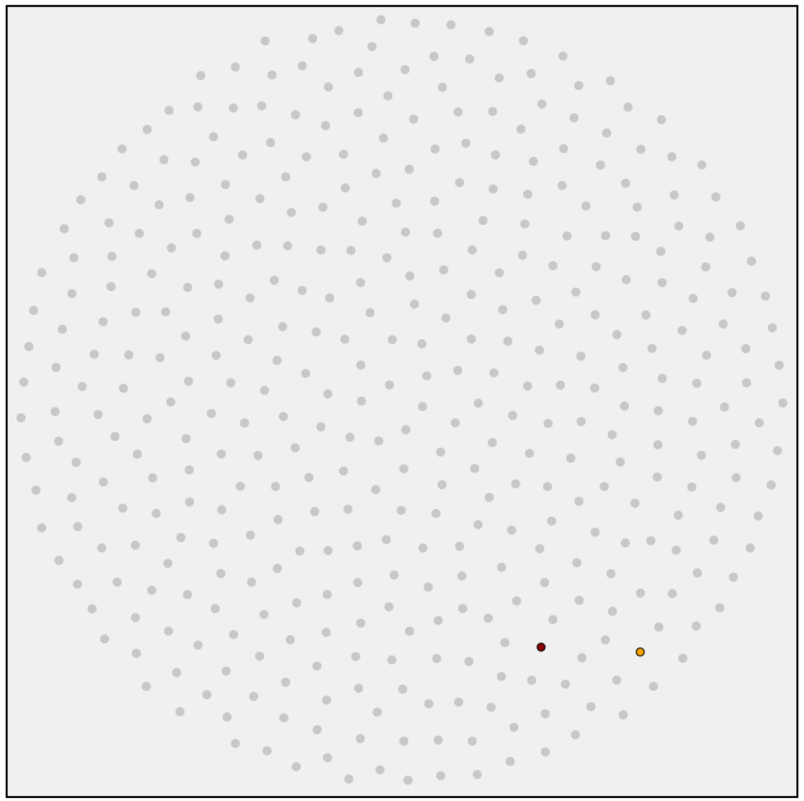
G(n, p) **G(n, m)**

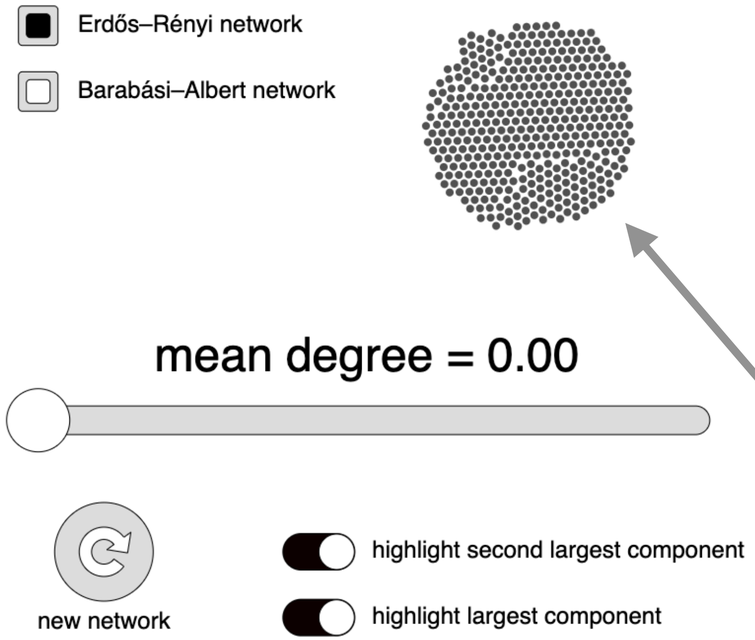
$$p * \frac{N * (N - 1)}{2} = m$$

pairs of nodes the graph has

Both are explanatory models, they ask: How do components emerge?

Let’s play





1. Set parameters as on the left.

2. Slowly move the slider from 0

3. Observe what happens to the blob, means degree and components

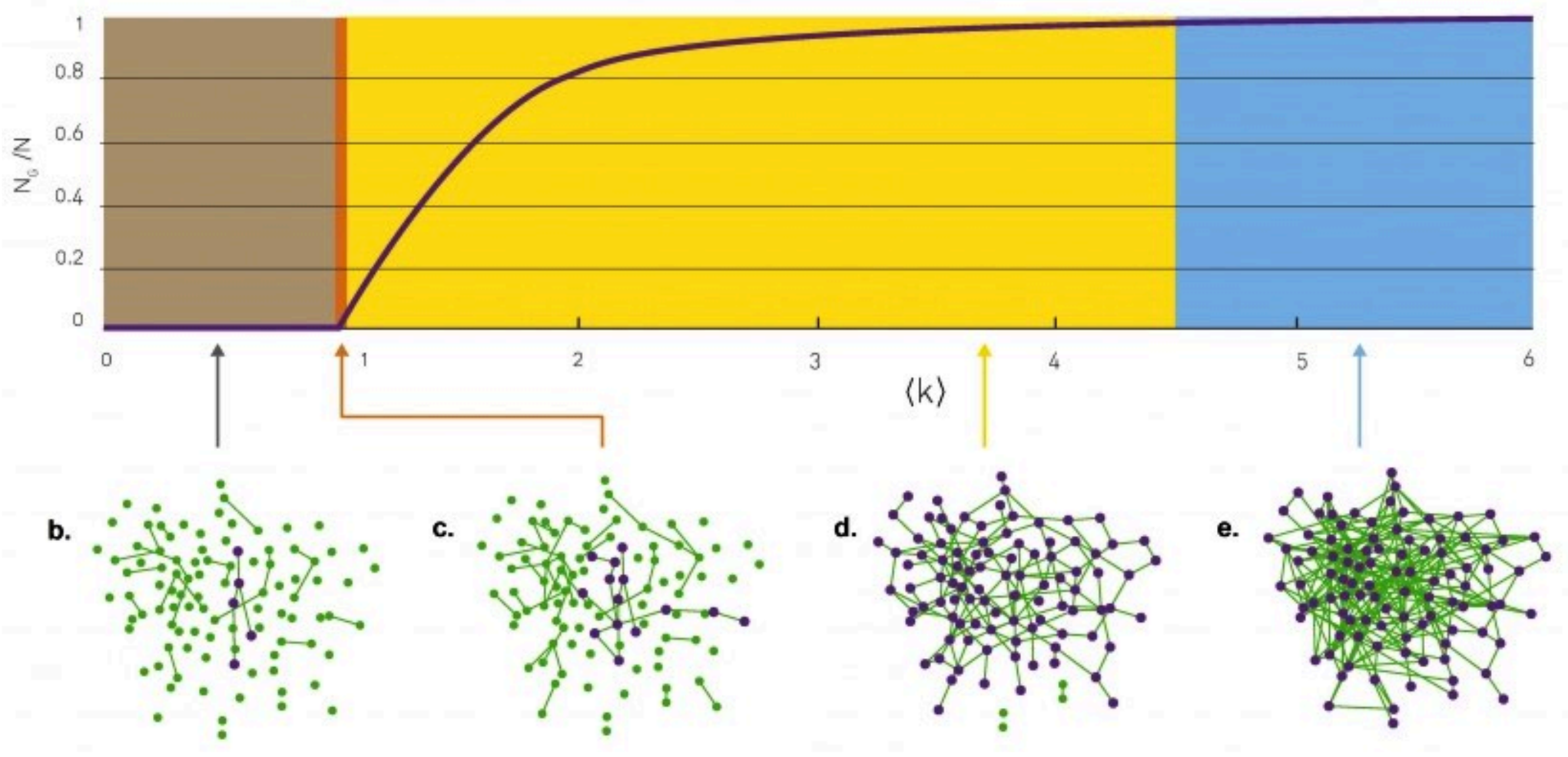
www.complexity-explorables.org —> The blob (Network Science)

ER model behaviour

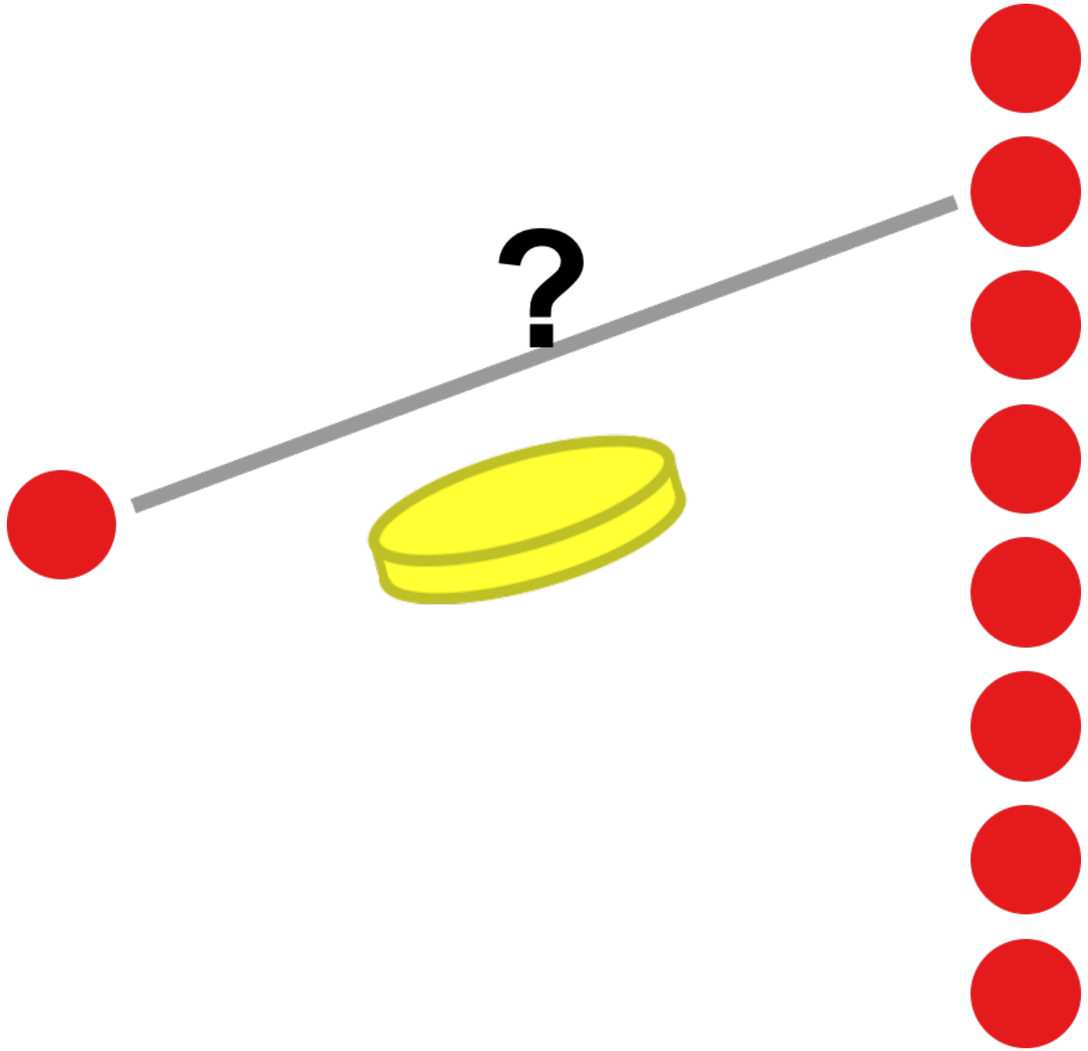
- Focus: **connected components**
  - For $p = 0 \Rightarrow$ **no links**: N components with one node each
  - For $p = 1 \Rightarrow$ **all links are there**: one component (complete network) with N nodes
- **Question:** what happens as we add links to the network?
  - **Naïve expectation:** the size of the largest component grows smoothly with the number of links
  - **Wrong:** there is an abrupt increase for a given value of the link probability p

Regimes in the ER Model

What happens when we fix n and increase p in a ER model?



Degree in the ER Model

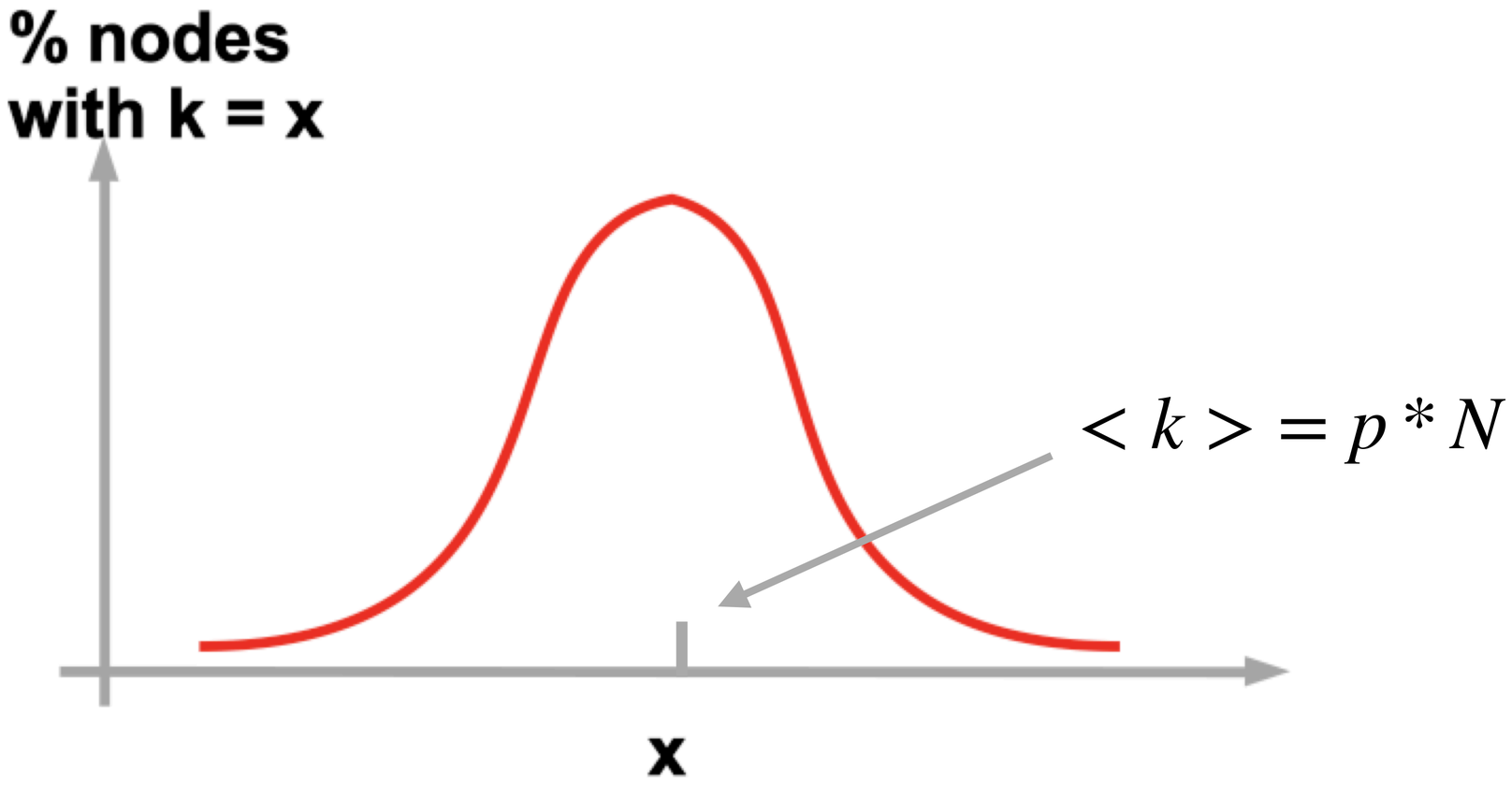


If p = 0.5

And there are 8 other nodes…

$k \sim 8 * 0.5 = 4 +/- \varepsilon$

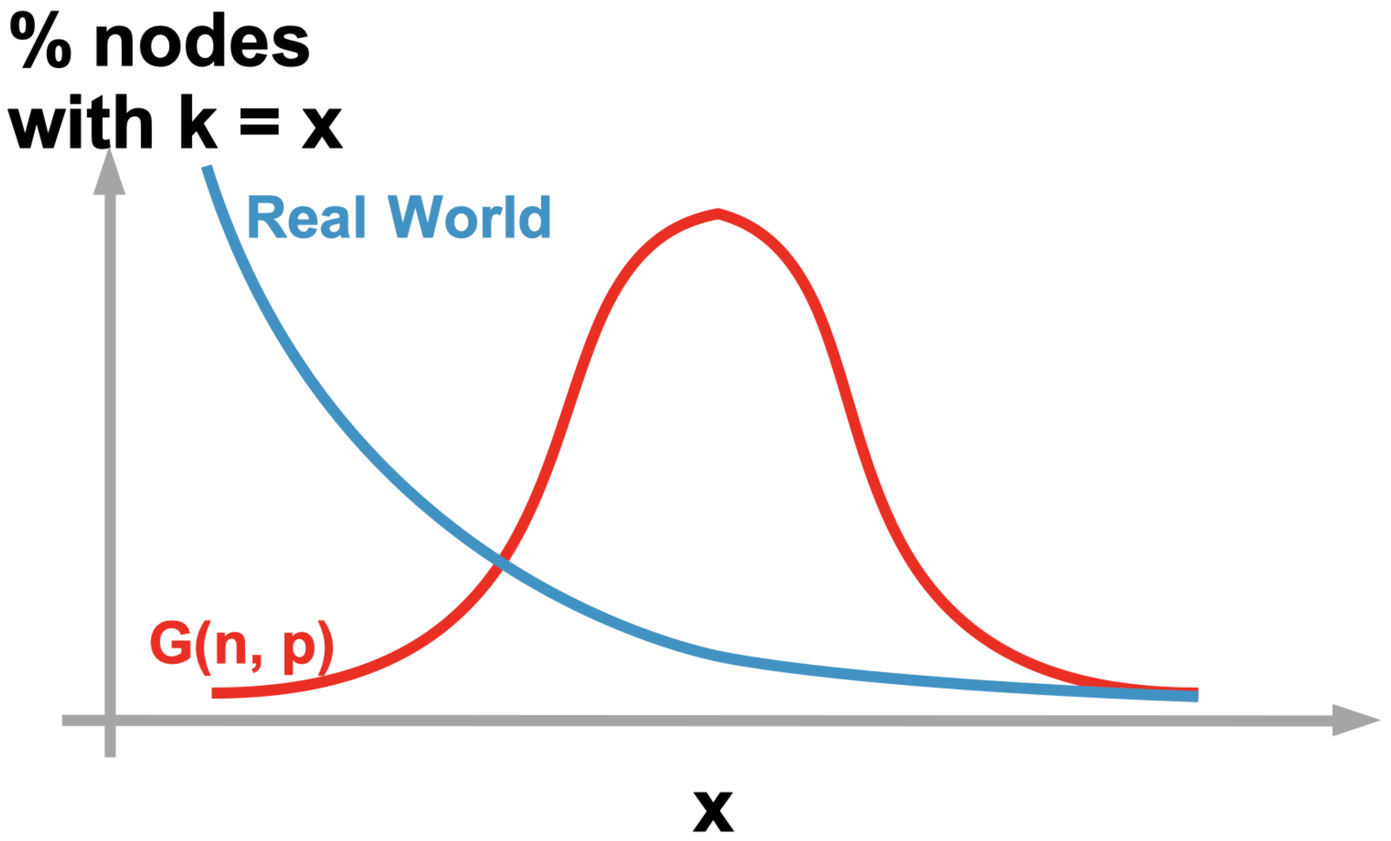
Degree in the ER Model



No hubs

Degree uniformly distributed

Degree in the ER Model

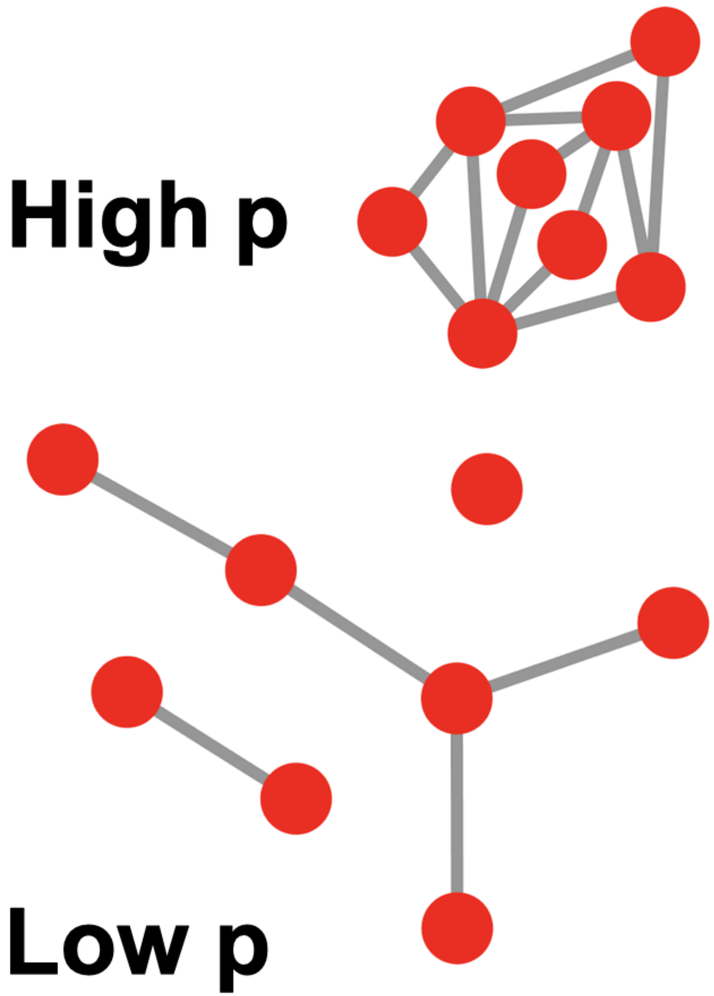


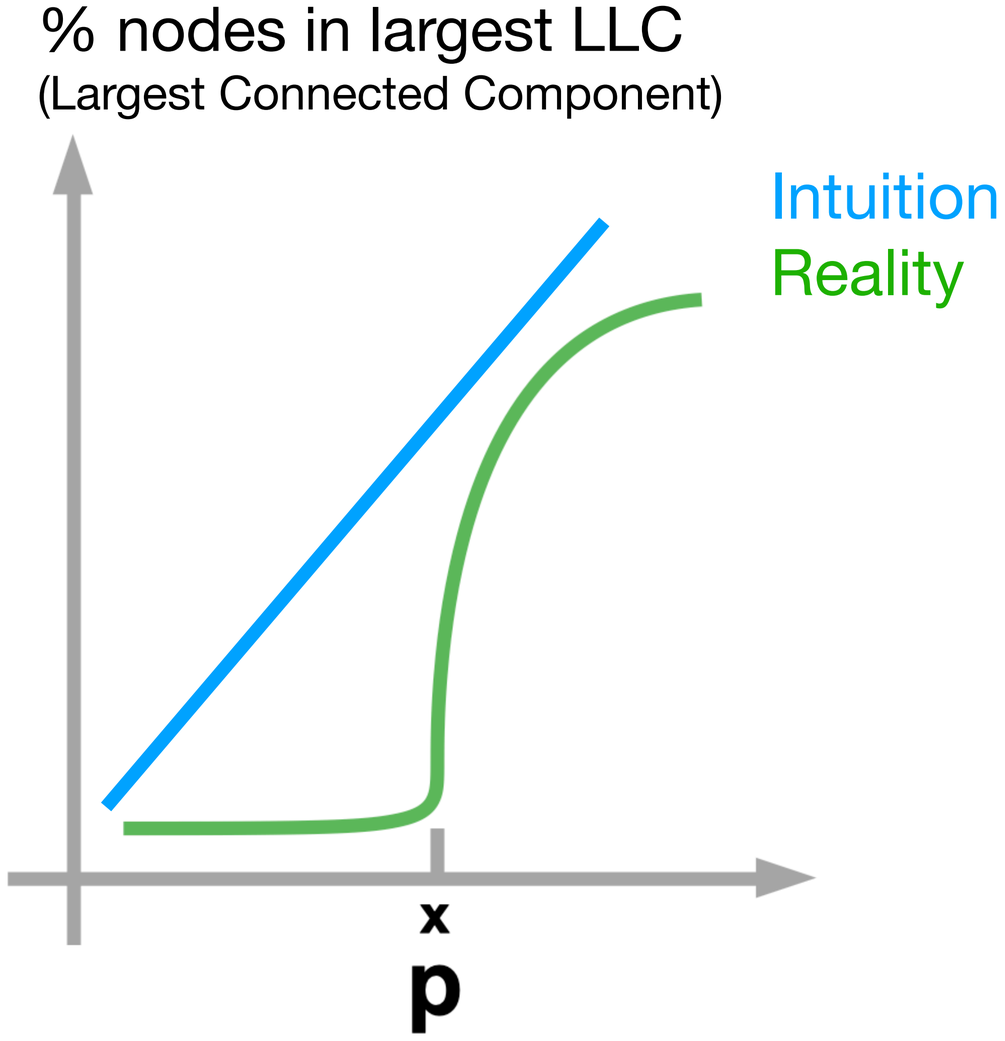
No hubs

Degree uniformly distributed

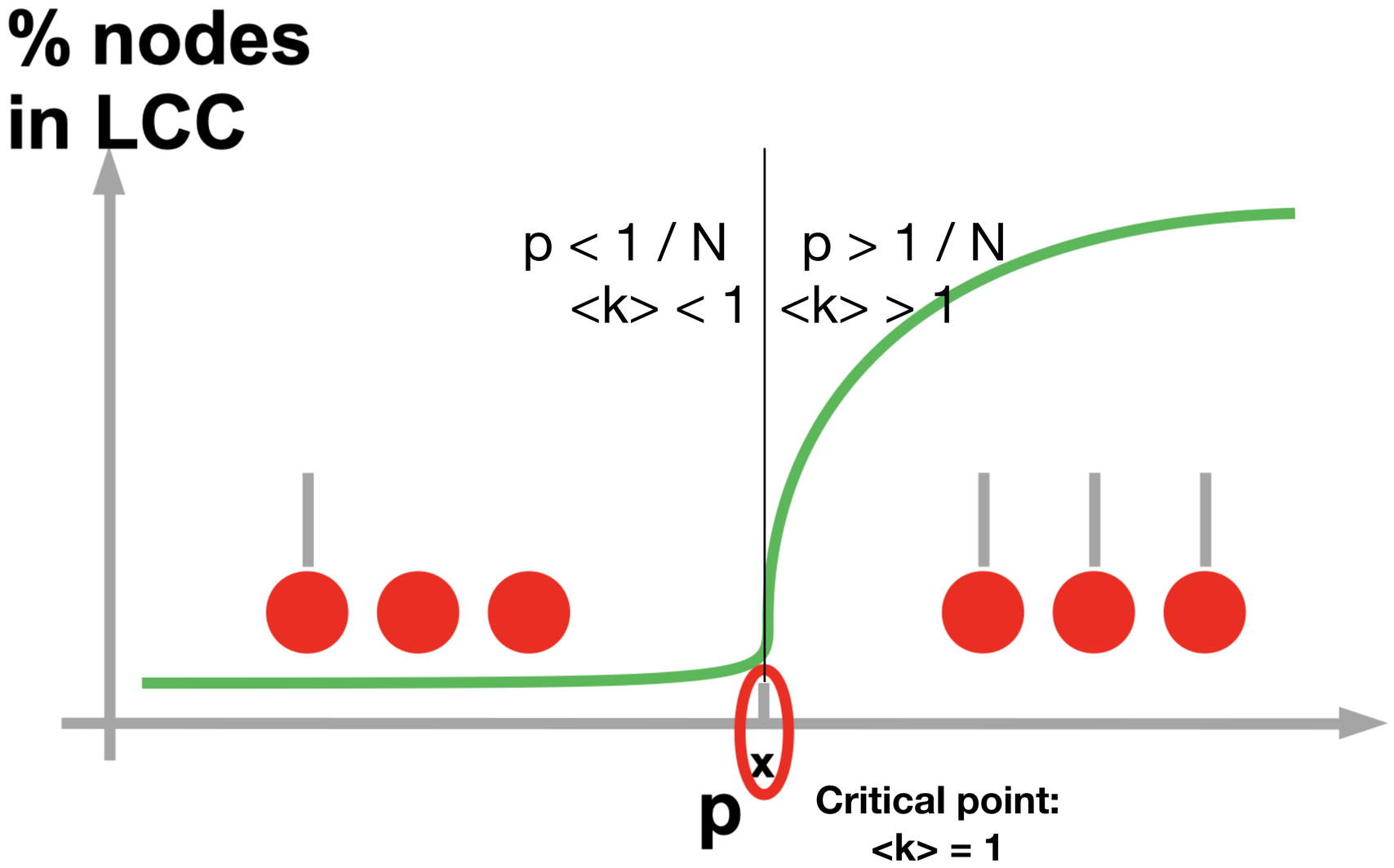
Nature has a very different story

Components in the ER Model

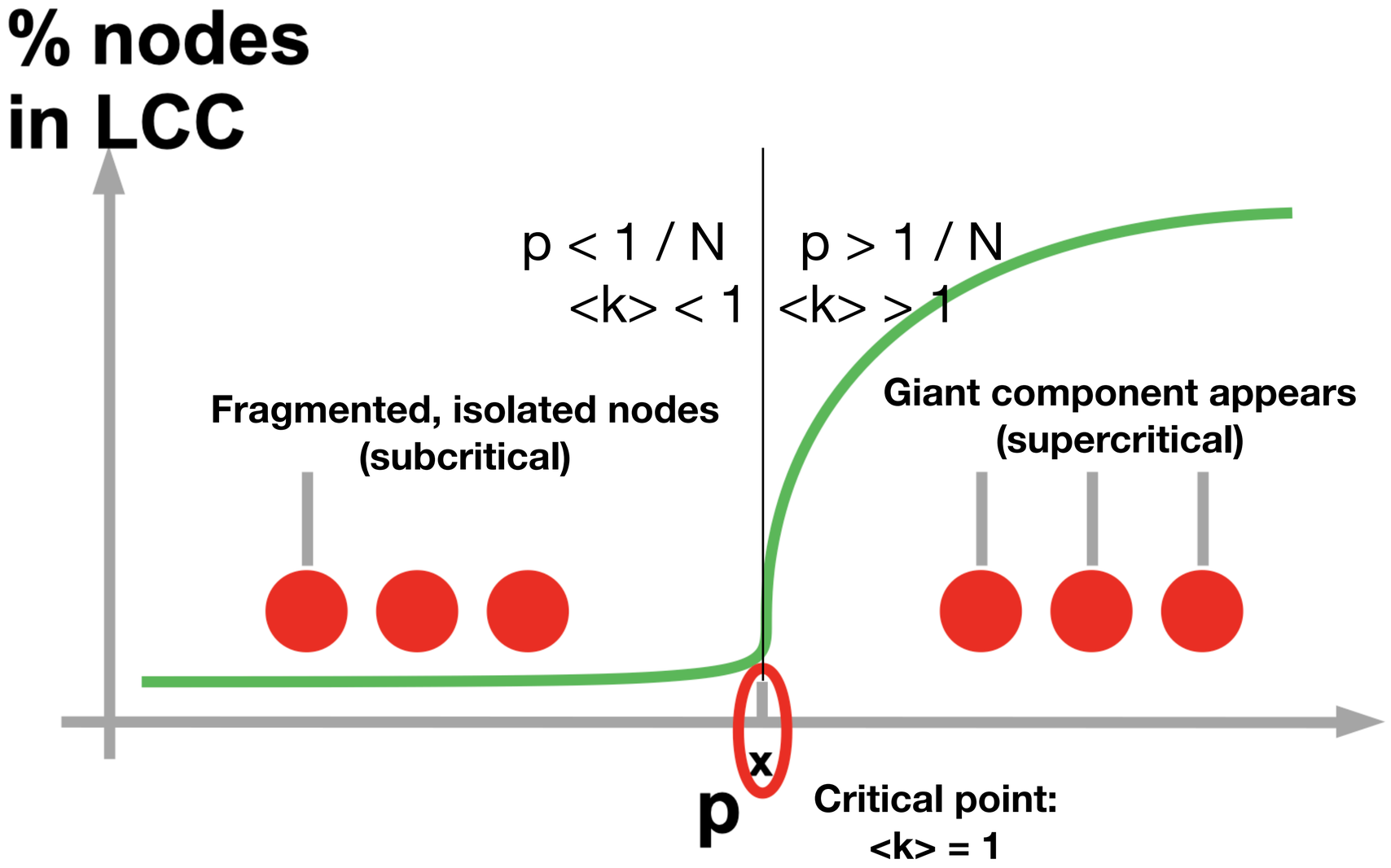




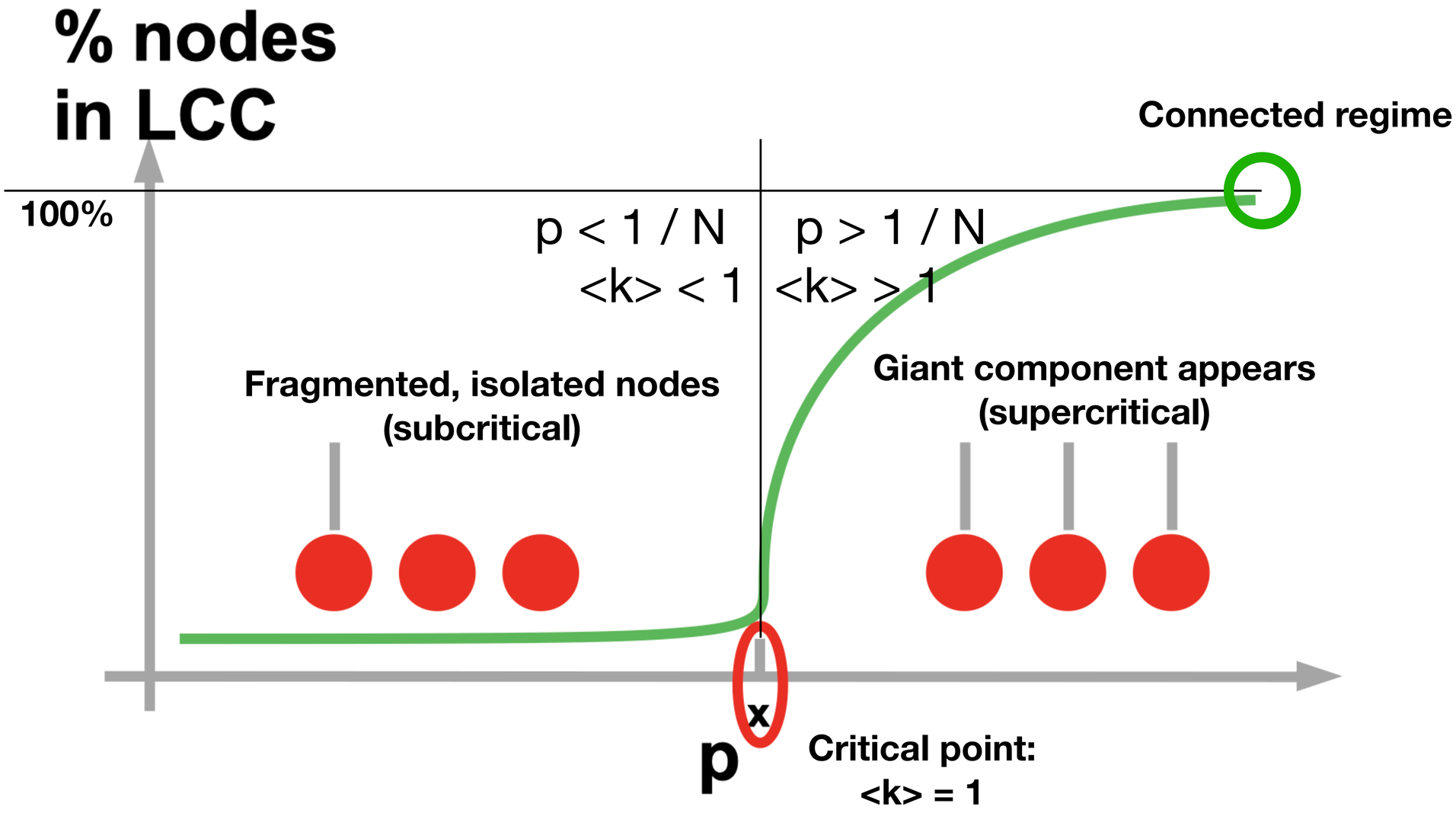
Components in the ER Model



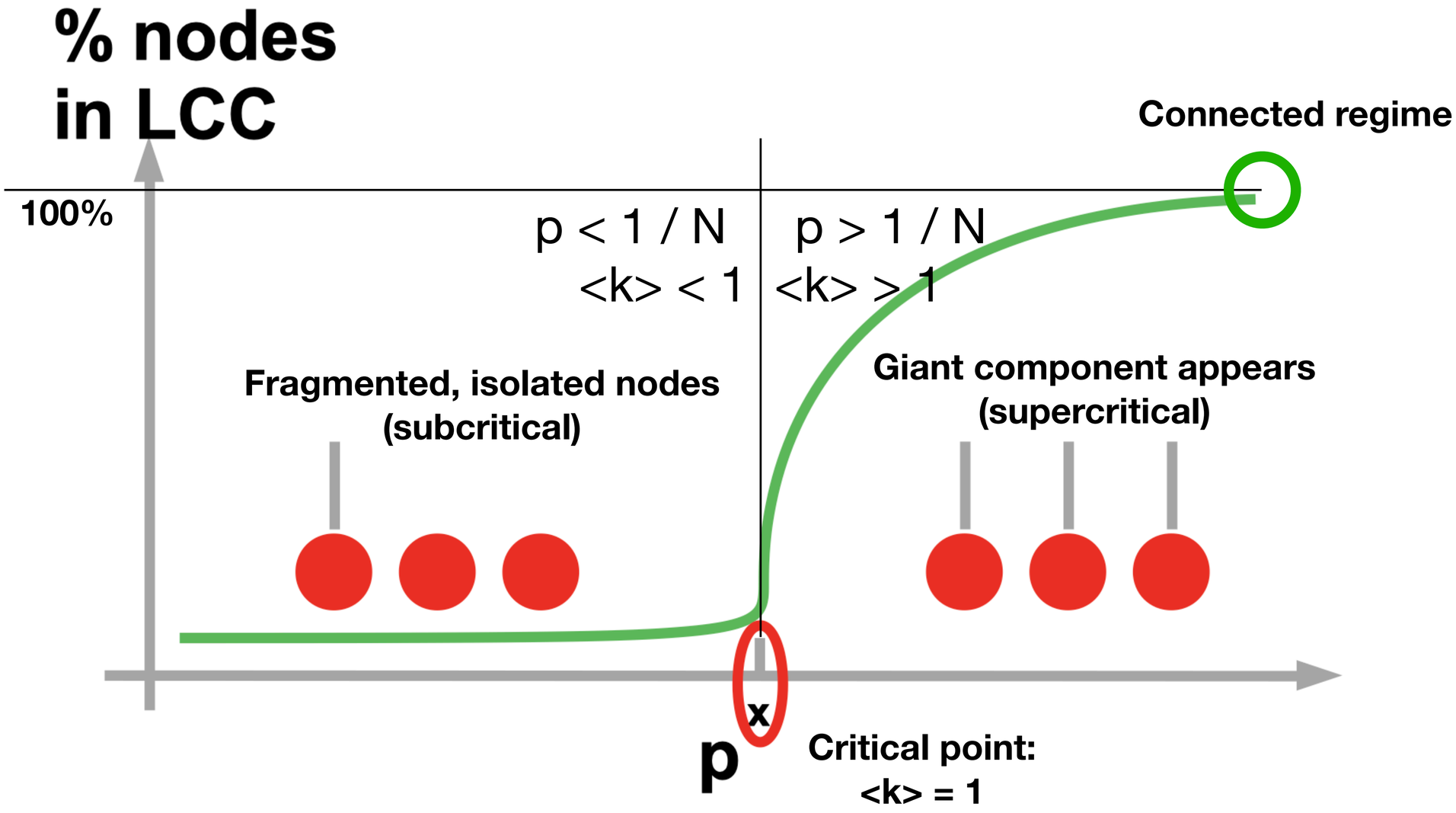
Components in the ER Model



Components in the ER Model



Components in the ER Model

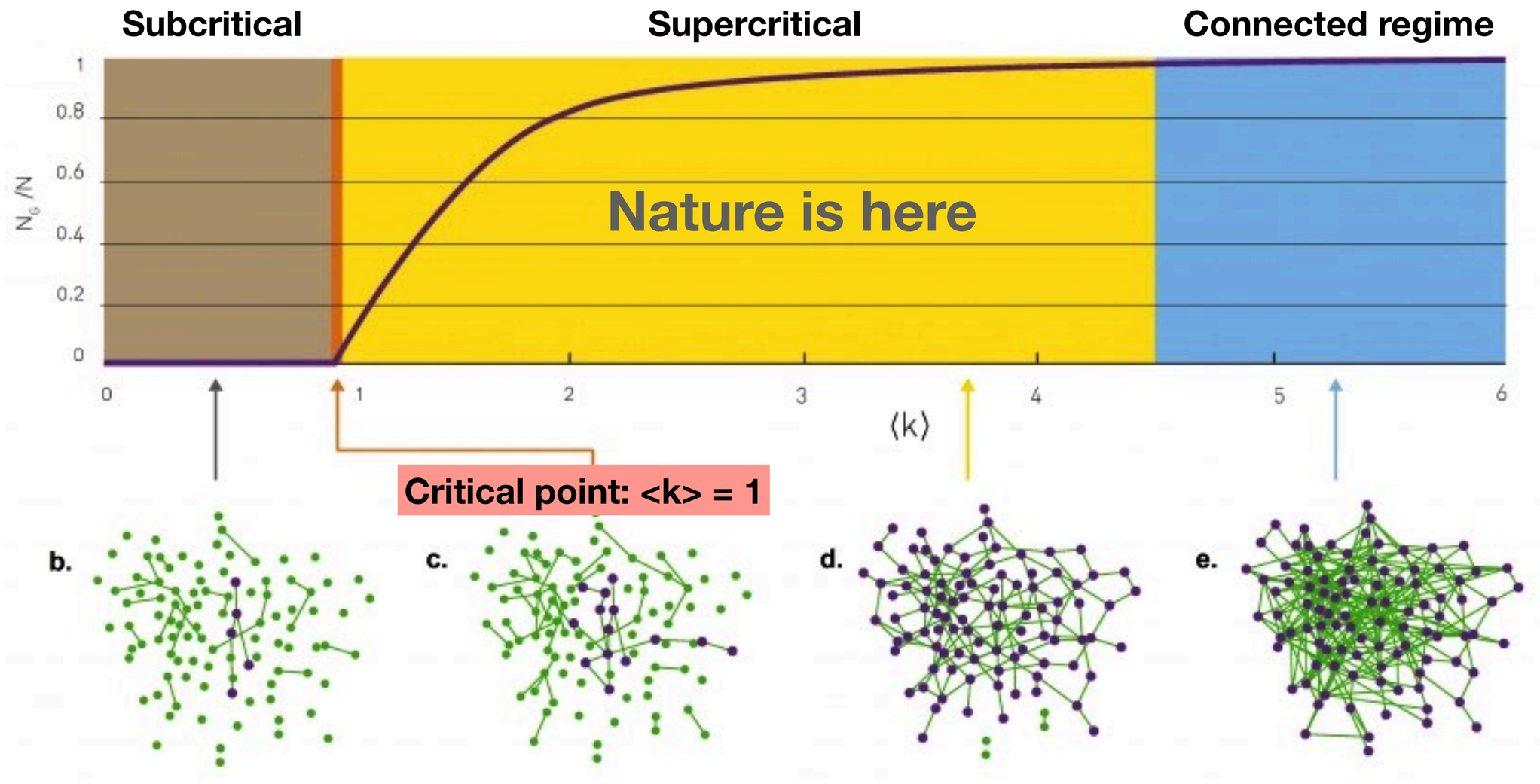


All nodes in LLC?

$\langle k \rangle > \ln(N)$

Regimes in the ER Model

Why do we care? It explains the formation of components and diameter in nature.



Clustering in the ER Model

We want the expected number of links L_i between the node's k_i neighbours.

As there are $k_i(k_i - 1)/2$ possible links between the k_i neighbours of node $\boldsymbol{i}$, the expected value is:

$$c_i = \frac{2\mathrm{L_i}}{\mathrm{k_i}(\mathrm{k_i}-1)}$$

Clustering in the ER Model

We want the expected number of links L_i between the node's k_i neighbours.

As there are $k_i(k_i - 1)/2$ possible links between the k_i neighbours of node $\boldsymbol{i}$, the expected value is:

$$c_i = \frac{2\mathrm{L_i}}{\mathrm{k_i(k_i-1)}} \qquad < L_i > = p\frac{k_i(k_i-1)}{2}$$

Clustering in the ER Model

We want the expected number of links L_i between the node's k_i neighbours.

As there are $k_i(k_i - 1)/2$ possible links between the k_i neighbours of node $\boldsymbol{i}$, the expected value is:

$$c_i = \frac{2\mathrm{L_i}}{\mathrm{k_i(k_i-1)}} \qquad < L_i > = p\frac{k_i(k_i-1)}{2}$$

The local clustering coefficient of a random network is:

$$c_i = \frac{2 < L_i >}{k_i(k_i-1)} = \frac{k_i(k_i-1)}{k_i(k_i-1)} = p = \frac{< k >}{N}$$

Clustering in the ER Model

We want the expected number of links L_i between the node's k_i neighbours.

As there are $k_i(k_i - 1)/2$ possible links between the k_i neighbours of node $\boldsymbol{i}$, the expected value is:

$$c_i = \frac{2\mathrm{L_i}}{\mathrm{k_i(k_i-1)}} \qquad < L_i > = p\frac{k_i(k_i-1)}{2}$$

The local clustering coefficient of a random network is:

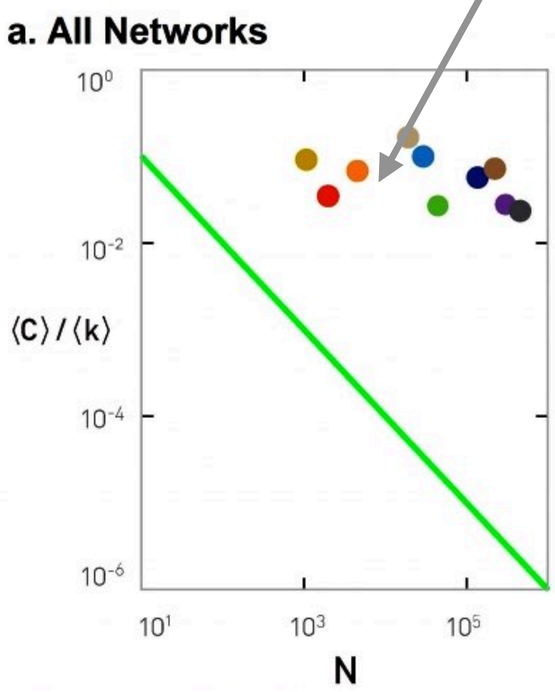
$$c_i = \frac{2 < L_i >}{k_i(k_i-1)} = \frac{k_i(k_i-1)}{k_i(k_i-1)} = p = \frac{< k >}{N}$$

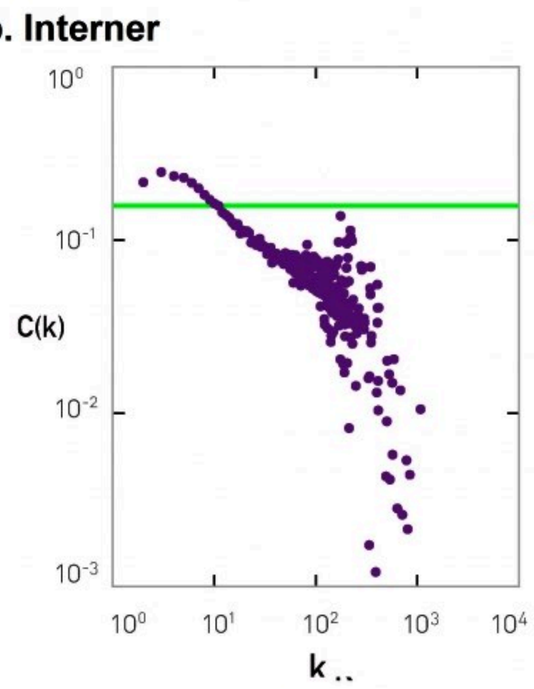
Much lower than the one of real world networks! ← **p** → Every node has the same (So it's also the average cc)

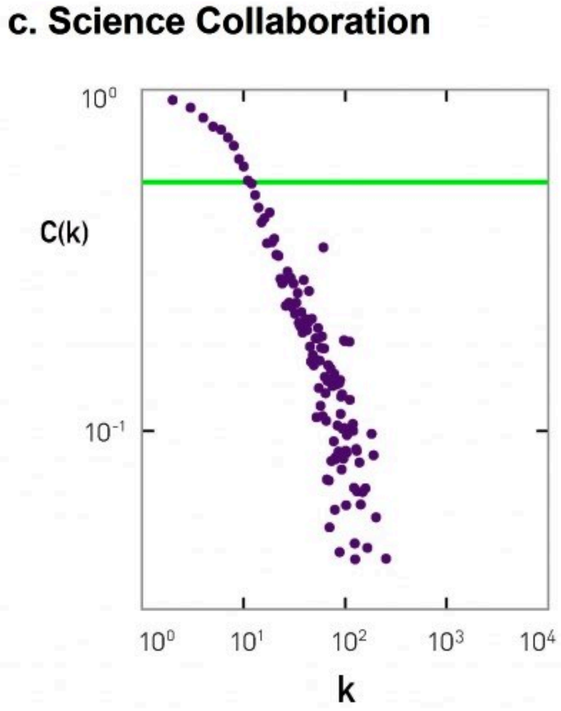
Clustering in the real networks

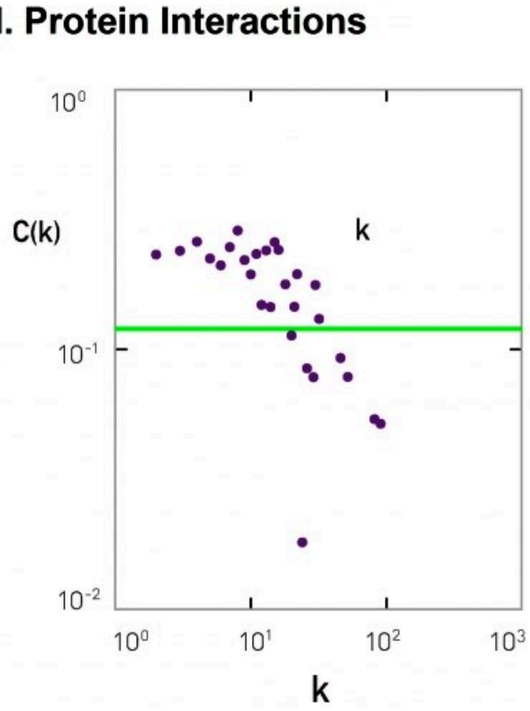
Real networks used in the book

Green line: prediction by ER model





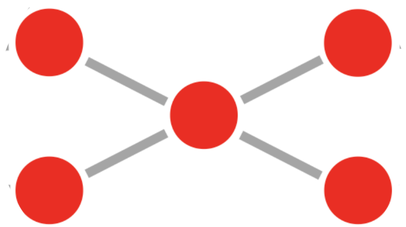




<C> independent of N

The ER model does not explain clustering formation

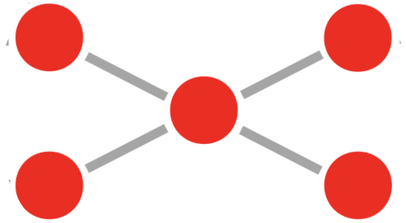
Avg. path length in the ER Model



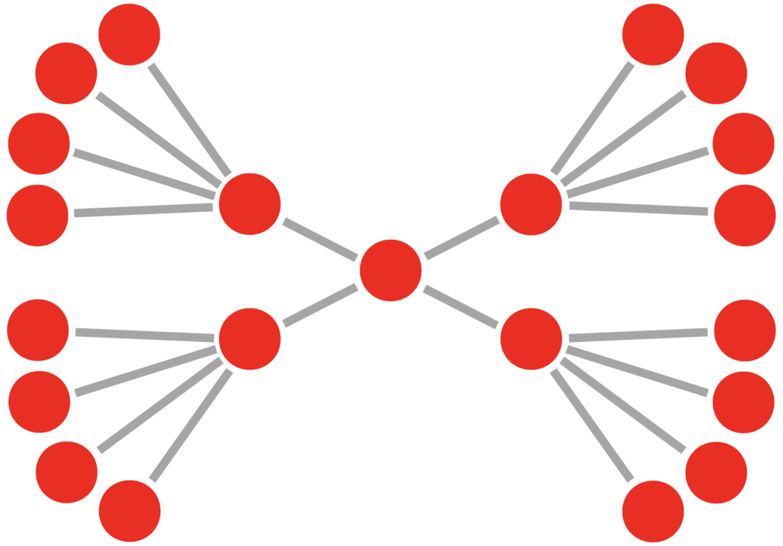
Number of nodes
distance 1 = <k>

Avg. path length in the ER Model

Random graphs tend to have a tree-like topology with almost constant node degrees k.



Number of nodes
distance 1 = <k>

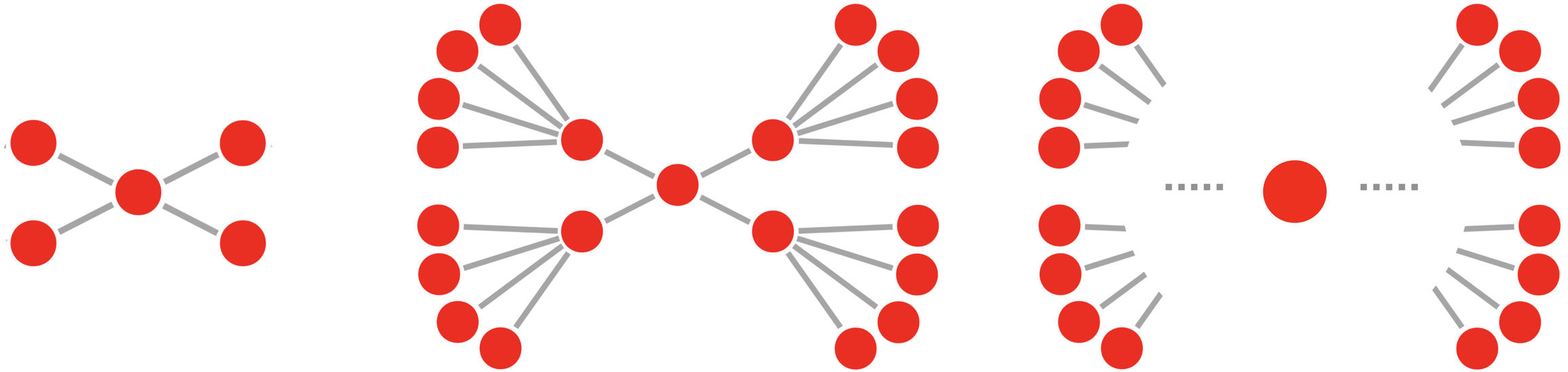


Number of nodes
distance 2 = <k^2>

Avg. path length in the ER Model

<k>^{d} = # of nodes ?

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



Number of nodes
distance 1 = <k>

Number of nodes
distance 2 = <k>^{2}

Number of nodes
distance d = <k>^{d}

Summary ER Model

- Links are placed at random, independently of each other
- Distances between pairs of nodes are short (small-world property): **good!**
- The average clustering coefficient is much lower than on real networks of the same size and average degree: **bad!**
- The nodes have approximately the same degree, there are no hubs: **bad!**
- **Conclusion:** the random network is not a good model of many real-world networks

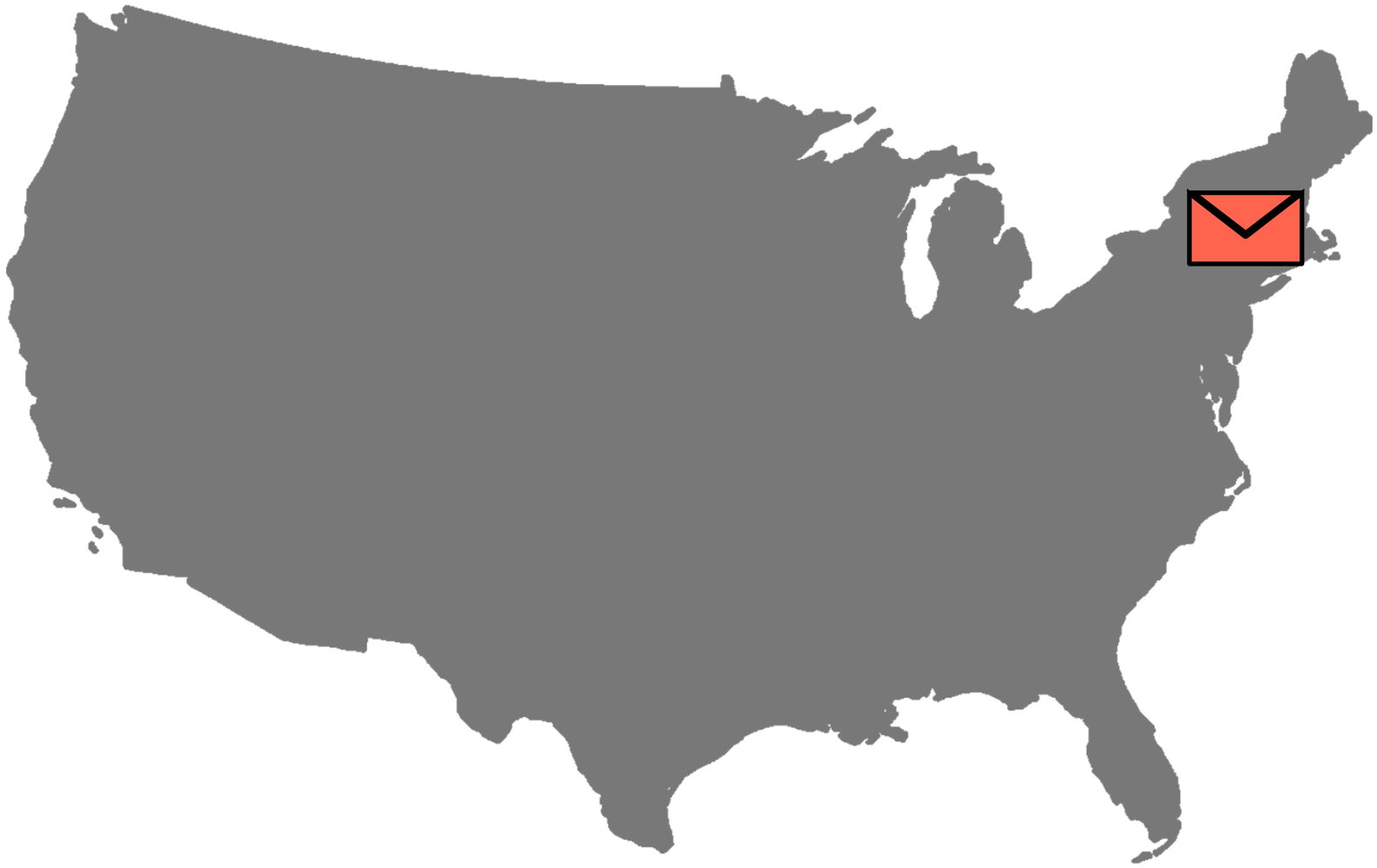
Small world

Stanley Milgram (1933-1984)

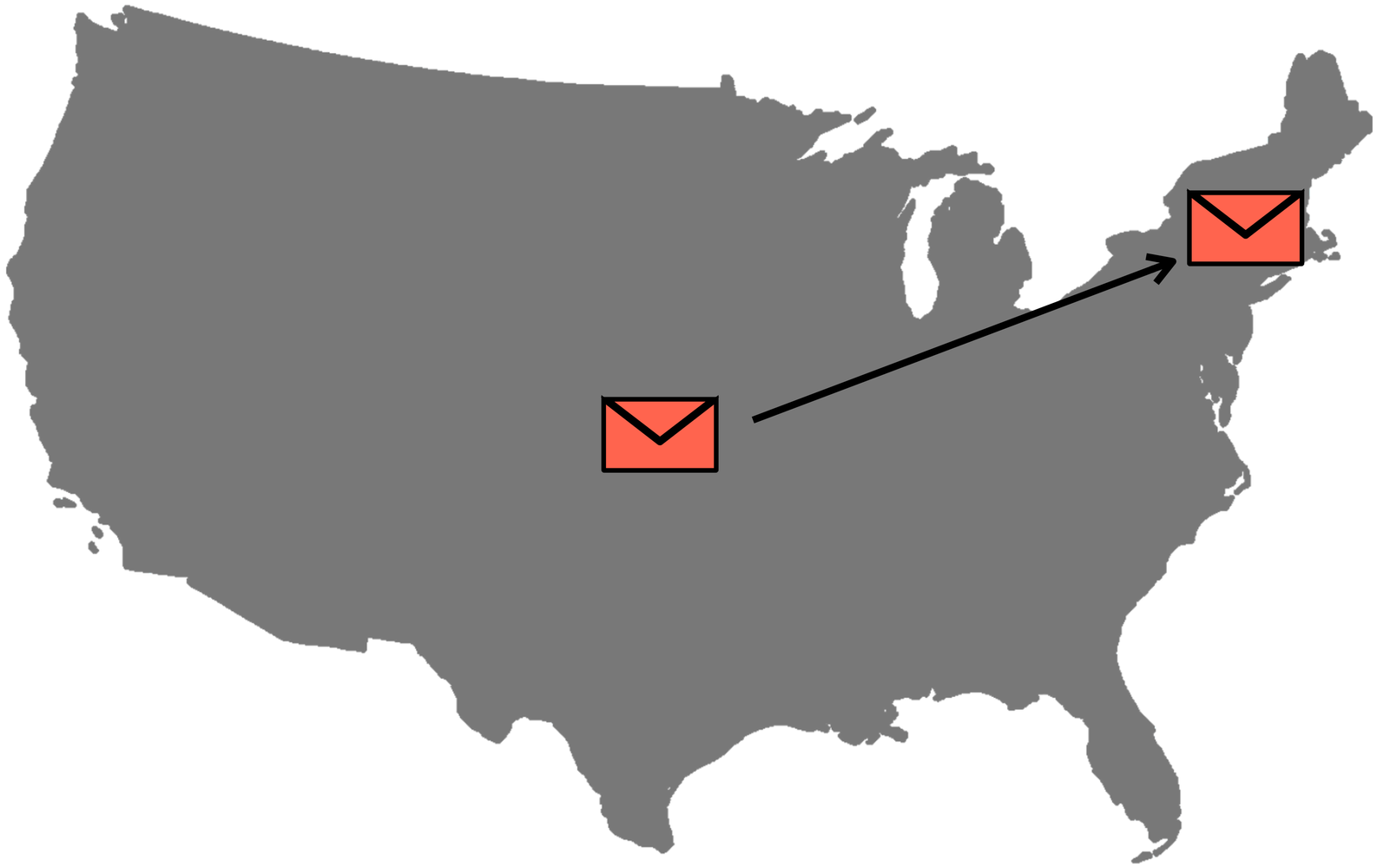
Which property of real-world networks is absent in ER networks?

Stanley Milgram
First small world experiment

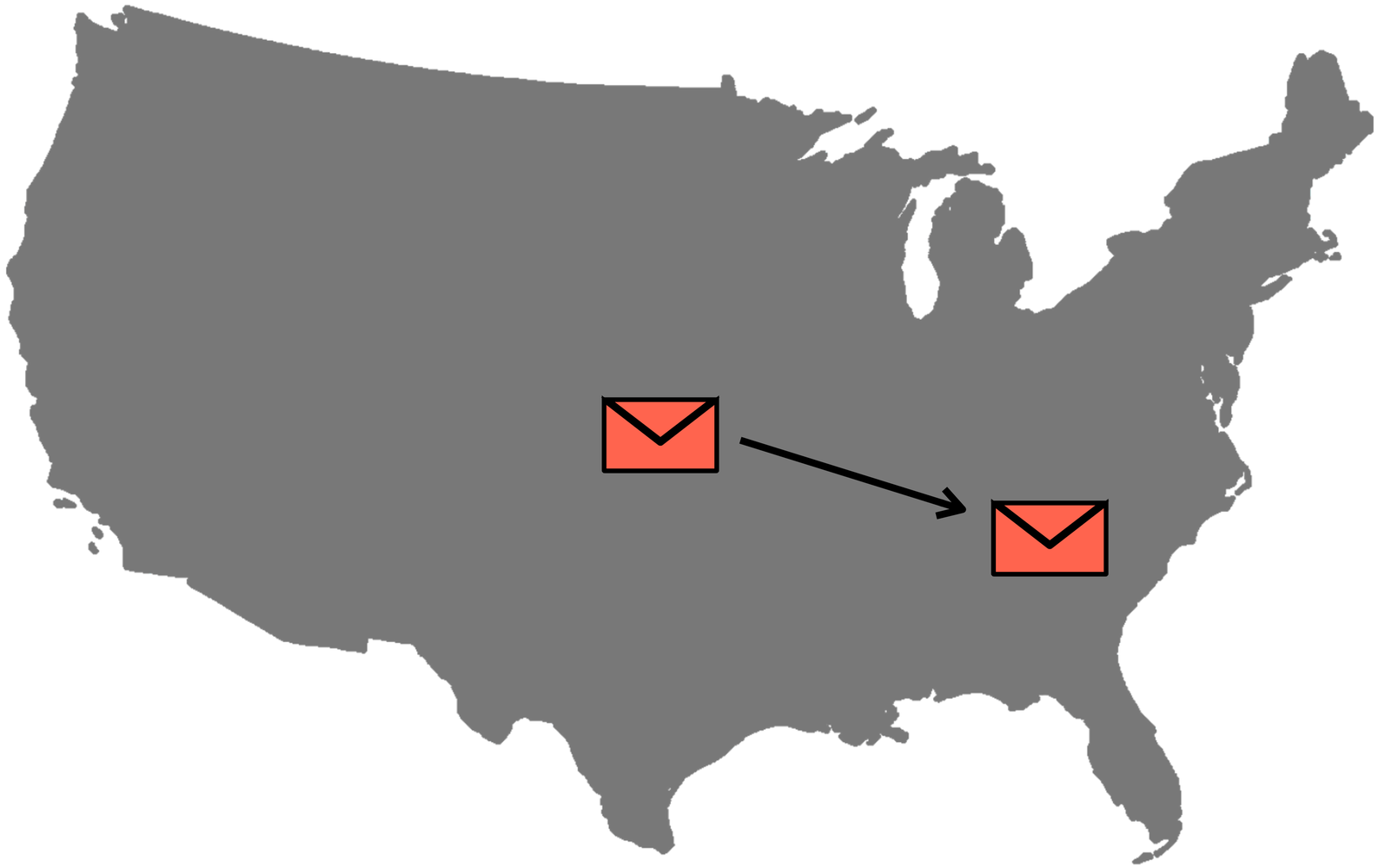
Stanley Milgram (1933-1984)



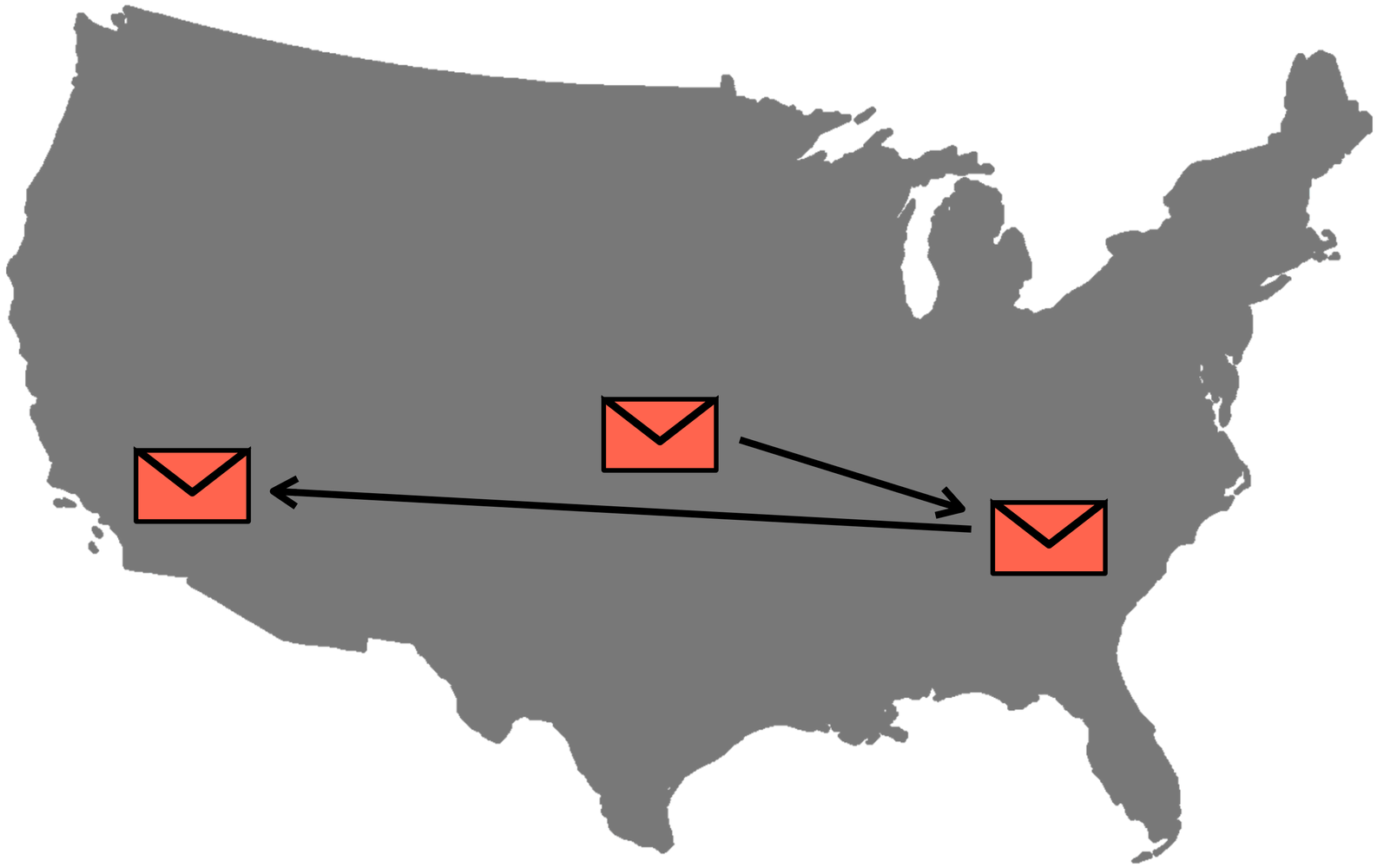
Stanley Milgram (1933-1984)



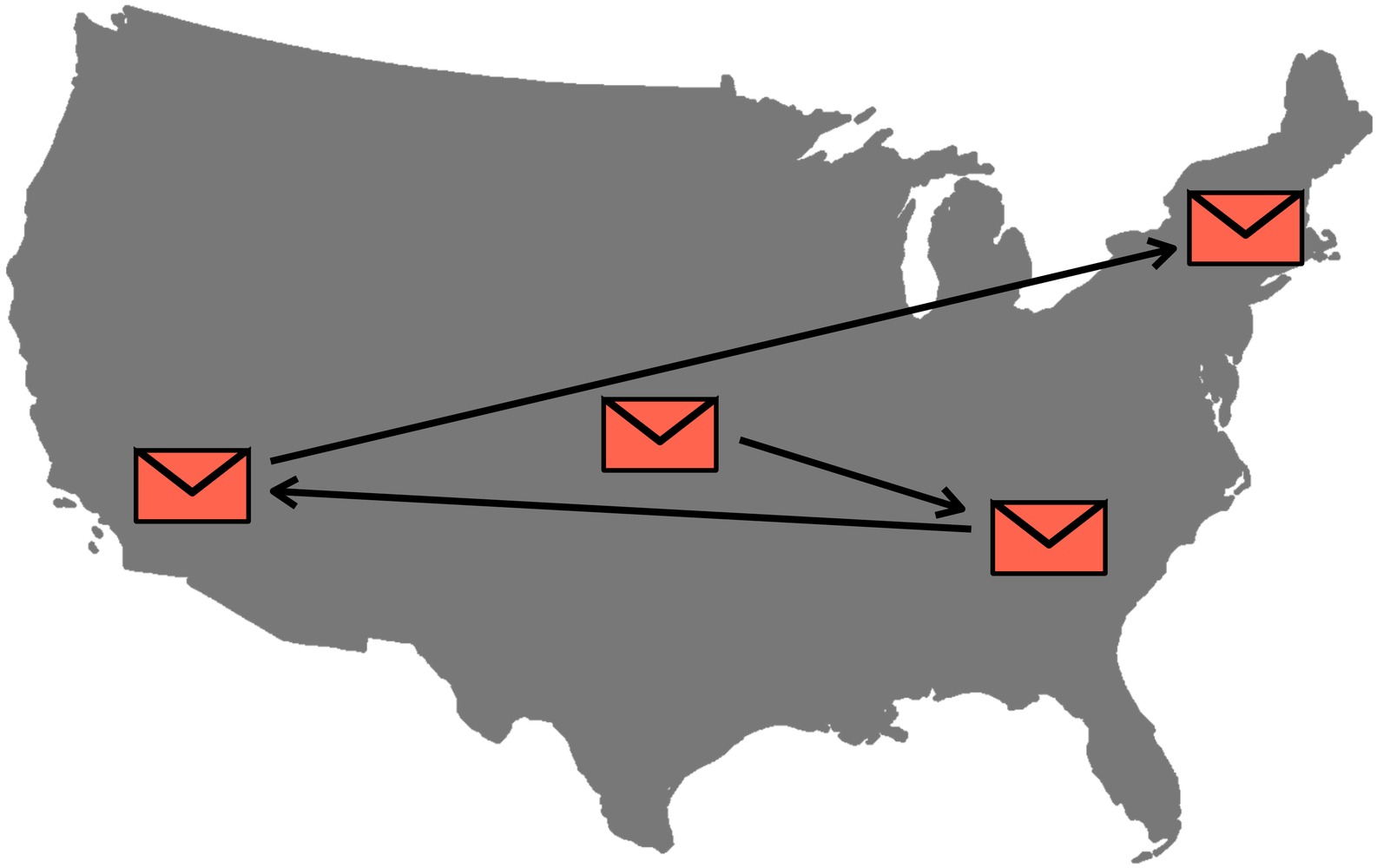
Stanley Milgram (1933-1984)



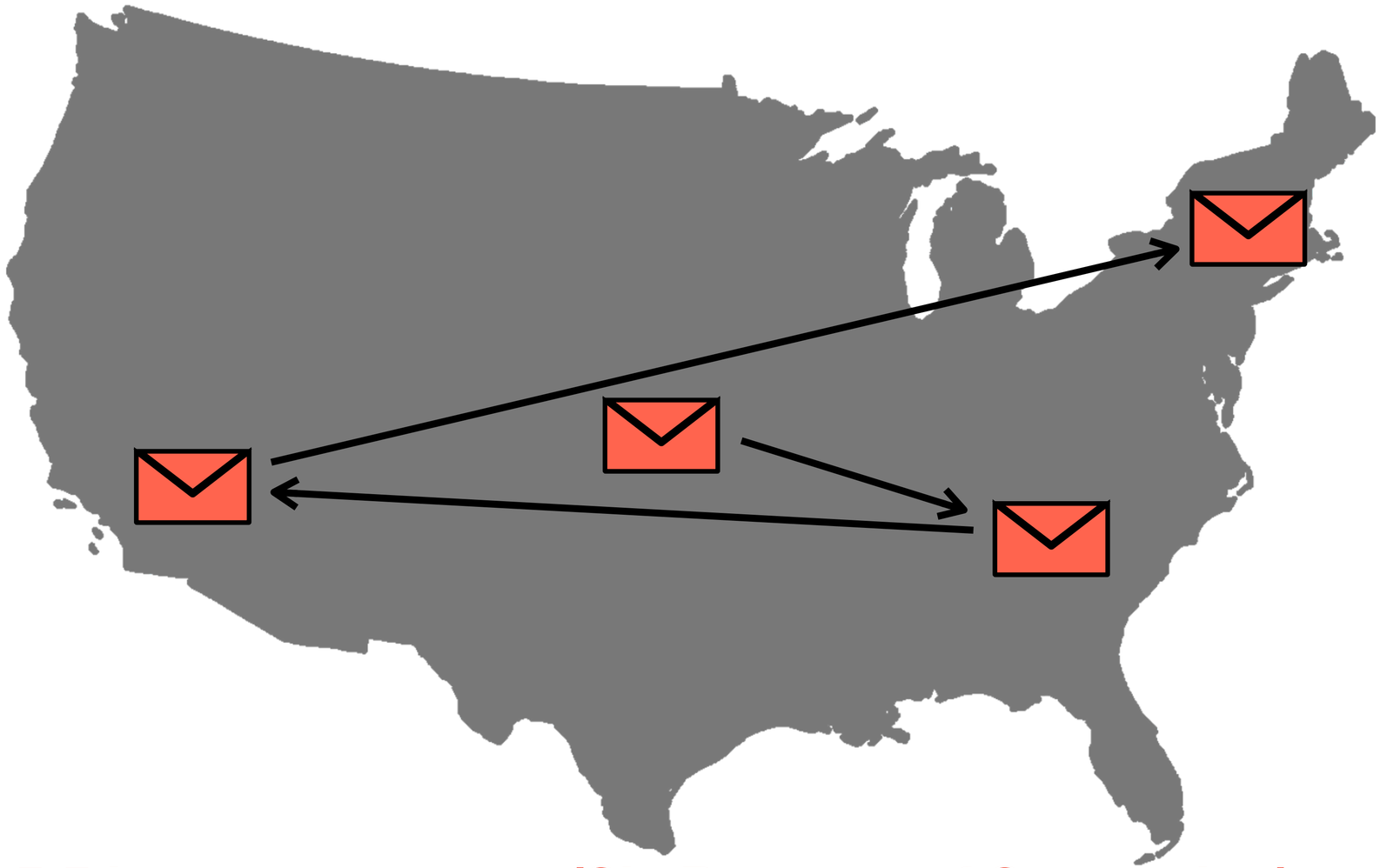
Stanley Milgram (1933-1984)



Stanley Milgram (1933-1984)

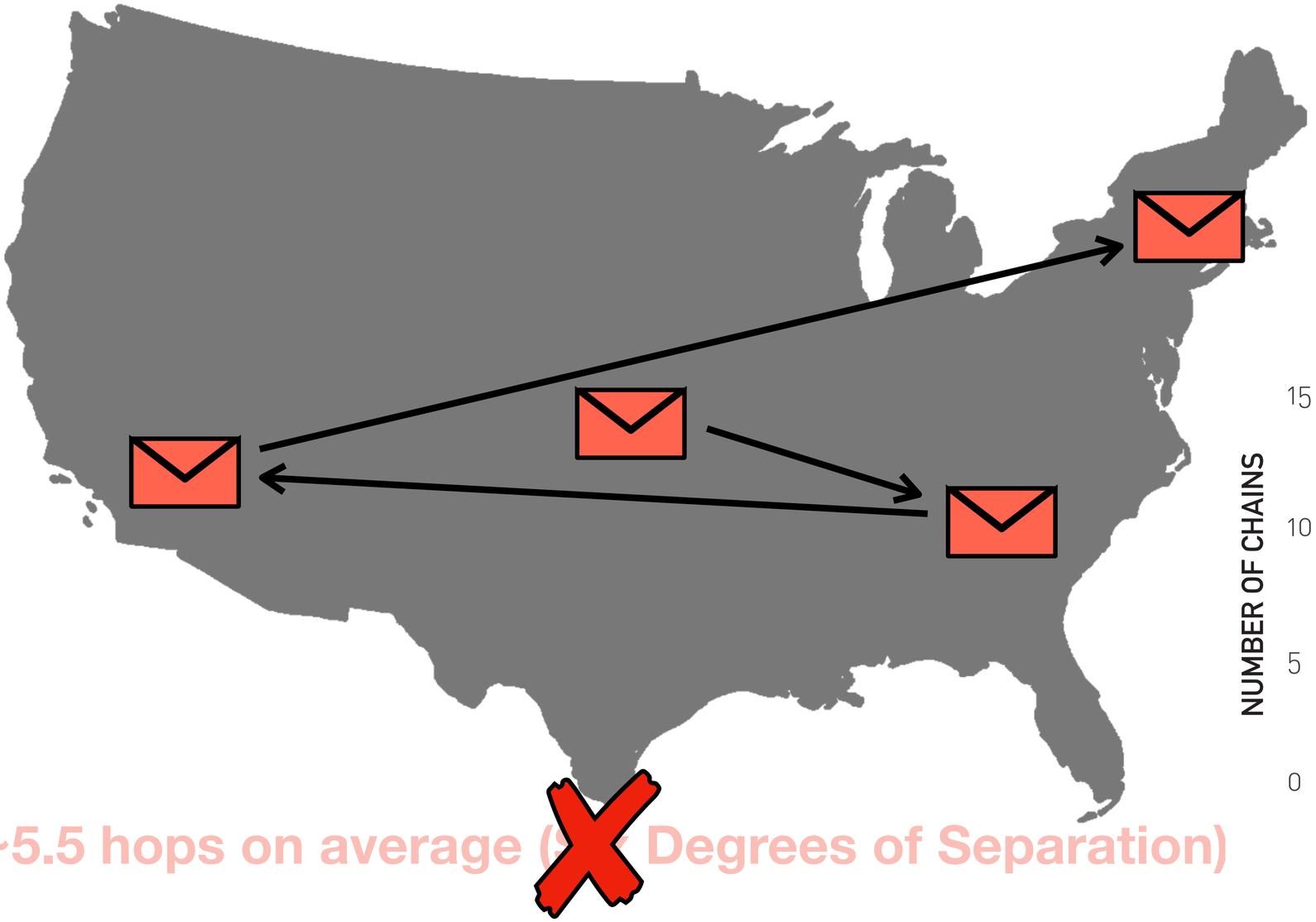


Stanley Milgram (1933-1984)

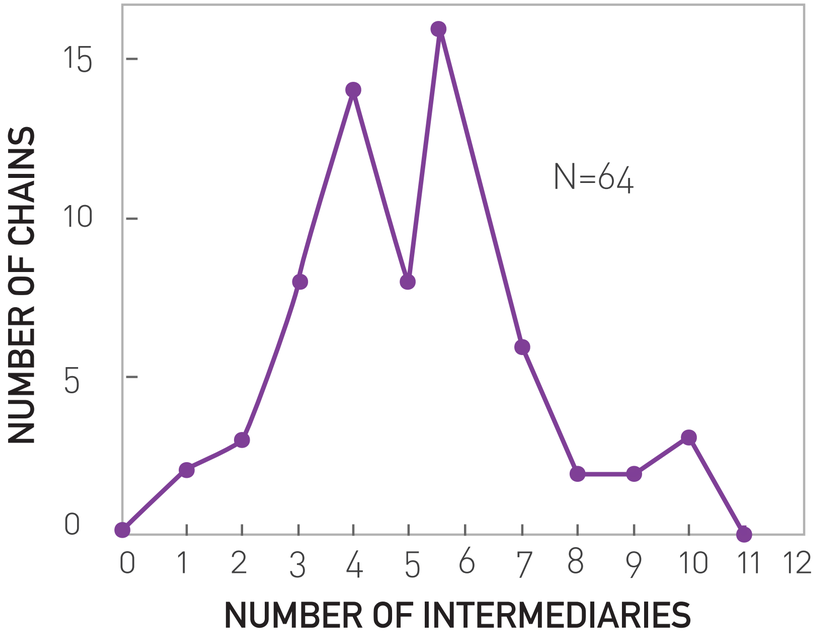


~5.5 hops on average (Six Degrees of Separation)

Stanley Milgram (1933-1984)

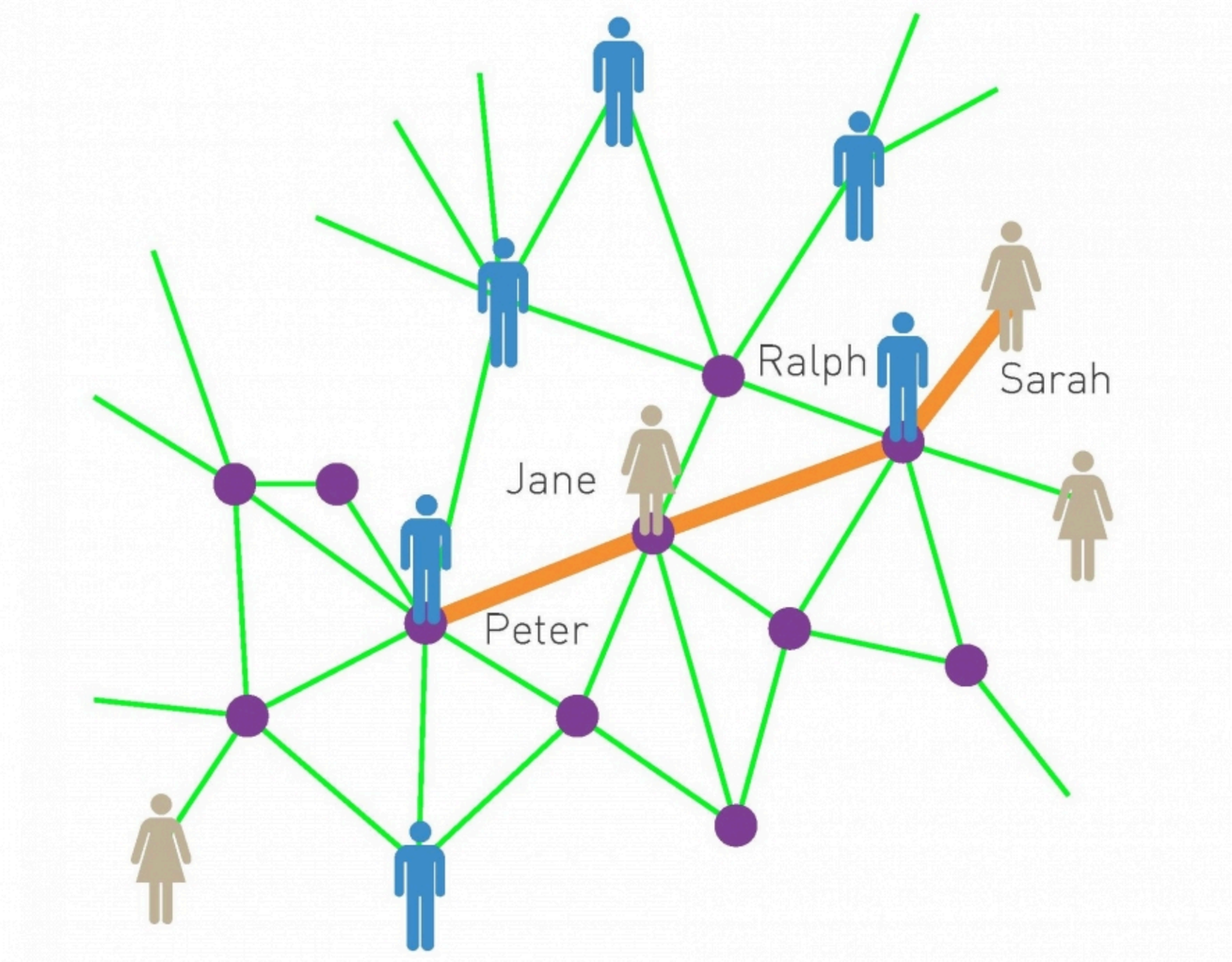


296 Sent…
64 Arrived!
< 22%



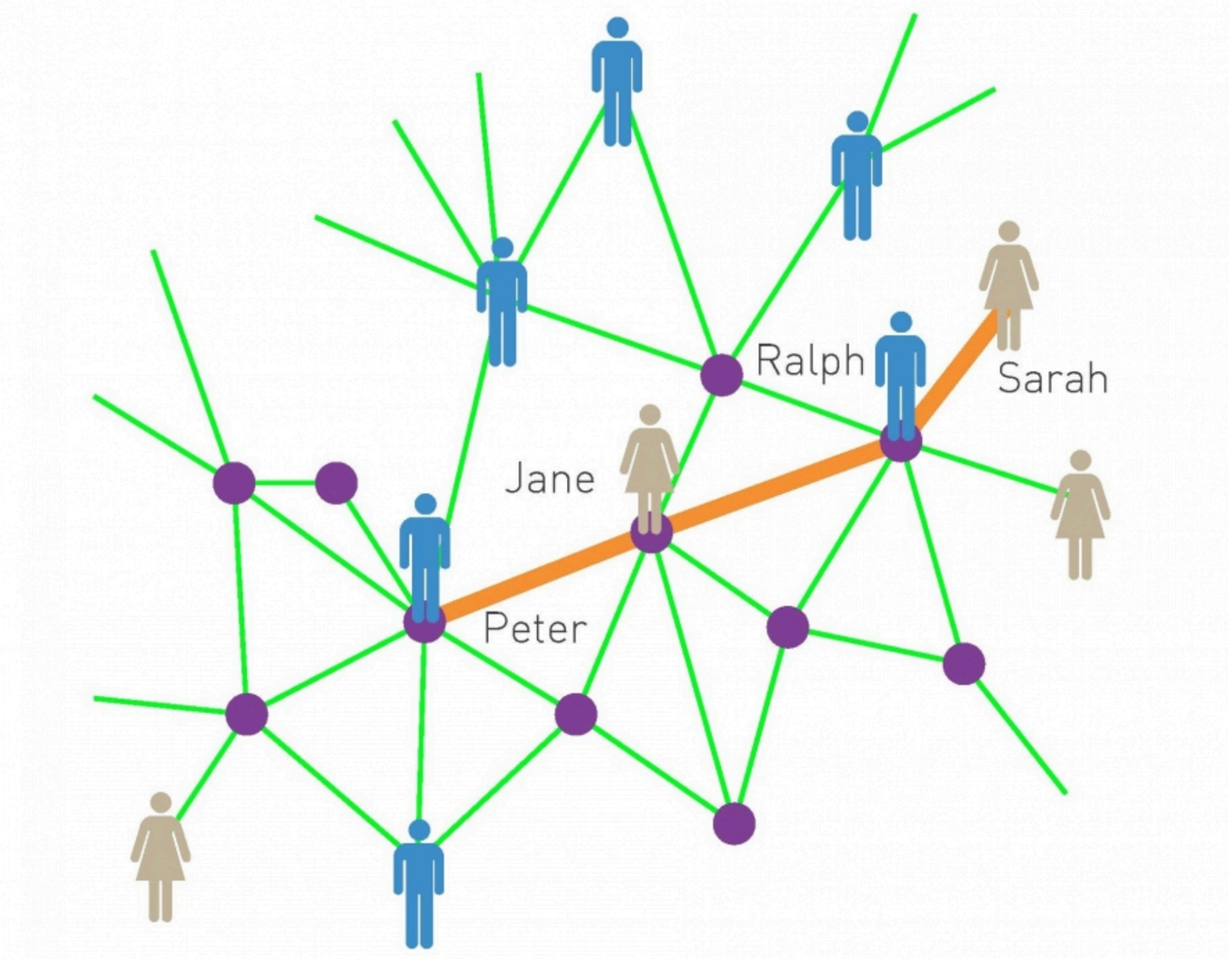
~5.5 hops on average (6 Degrees of Separation)

Small world property



If you choose any two individuals anywhere on Earth, you will find a path of at most six acquaintances between them.

Small world property



In a random network:
Number of nodes distance 1 $\langle k \rangle$
Number of nodes distance 2 $\langle k \rangle^2$
Number of nodes distance 3 $\langle k \rangle^3$
...
Number of nodes at distance $d \sim \langle k \rangle^d$

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

Small world property: The average path length $\langle d \rangle$ in a network grows like the logarithm of the network size N.

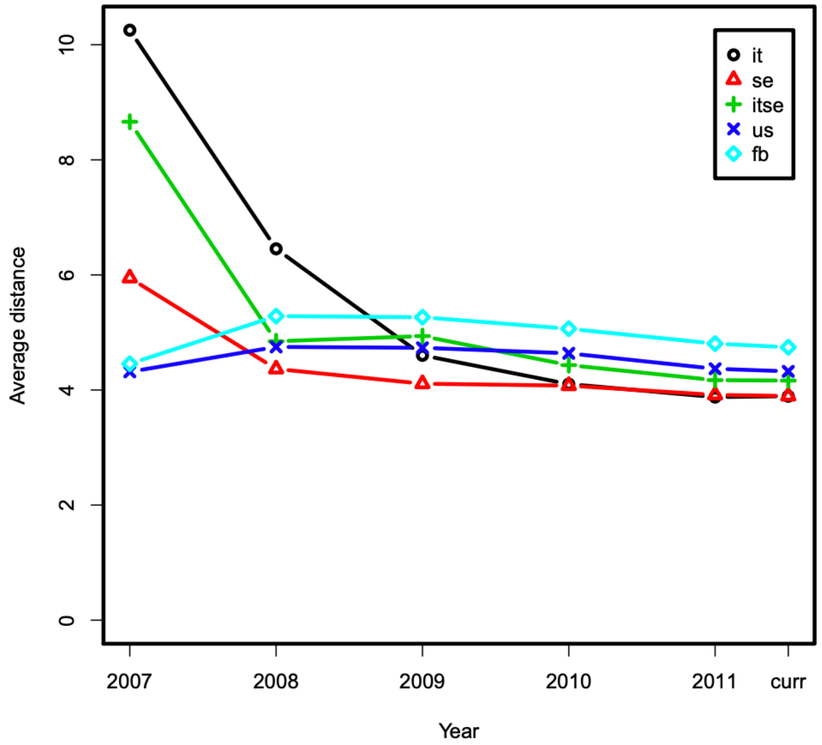
If you choose any two individuals anywhere on Earth, you will find a path of at most six acquaintances between them.

Degrees of separation - Small world

	it	se	itse	us	fb
2007	1.31	3.90	1.50	119.61	99.50
2008	5.88	46.09	36.00	106.05	76.15
2009	50.82	69.60	55.91	111.78	88.68
2010	122.92	100.85	118.54	128.95	113.00
2011	198.20	140.55	187.48	188.30	169.03
current	226.03	154.54	213.30	213.76	190.44

Average degree increasing

Table 4: Average degree of the datasets.



Path length decreasing

Four Degrees of Separation

Lars Backstrom
Facebook
lars@fb.com

Paolo Boldi
Univ. degli Studi di Milano
boldi@dsi.unimi.it

Marco Rosa
Univ. degli Studi di Milano
marco.rosa@unimi.it

Johan Ugander
Facebook
jugander@fb.com

Sebastiano Vigna*
Univ. degli Studi di Milano
vigna@acm.org

ABSTRACT

Frigyes Karinthy, in his 1929 short story "Láncszemek" (in English, "Chains") suggested that any two persons are distanced by at most six friendship links.[1] Stanley Milgram in his famous experiments challenged people to route postcards to a fixed recipient by passing them only through direct acquaintances. Milgram found that the average number of intermediaries on the path of the postcards lay between 4.4 and 5.7, depending on the sample of people chosen. We report the results of the first world-scale social-network graph-distance computations, using the entire Facebook network of active users (≈ 721 million users, ≈ 69 billion friendship links). The average distance we observe is 4.74, corresponding to 3.74 intermediaries or "degrees of separation", prompting the title of this paper. More generally, we study the distance distribution of Facebook and of some interesting geographic subgraphs, looking also at their evolution over time. The networks we are able to explore are almost two orders of magnitude larger than those analysed in the previous literature. We report detailed statistical metadata showing that our measurements (which rely on probabilistic algorithms) are very accurate.

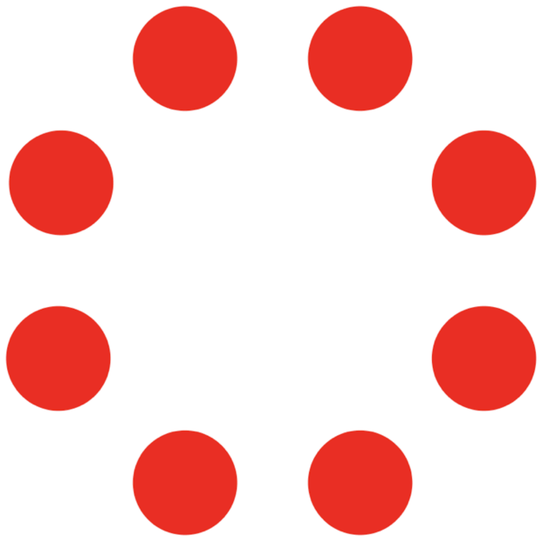
At the 20th World–Wide Web Conference, in Hyderabad, India, one of the authors (Sebastiano) presented a new tool for studying the distance distribution of very large graphs: HyperANF [3]. Building on previous graph compression work [4] and on the idea of diffusive computation pioneered in [19], the new tool made it possible to accurately study the distance distribution of graphs orders of magnitude larger than what was previously possible.

One of the goals in studying the distance distribution is the identification of interesting statistical parameters that can be used to tell proper social networks from other complex networks, such as web graphs. More generally, the distance distribution is one interesting *global* feature that makes it possible to reject probabilistic models even when they match local features such as the in-degree distribution. In particular, earlier work [3] had shown that the *spid*[2], which measures the *dispersion* of the distance distribution, appeared to be smaller than 1 (underdispersion) for social networks, but larger than one (overdispersion) for web graphs. Hence, during the talk, one of the main open questions was "What is the spid of Facebook?".

Proceedings of the 4th Annual ACM Web Science Conference. 2012.

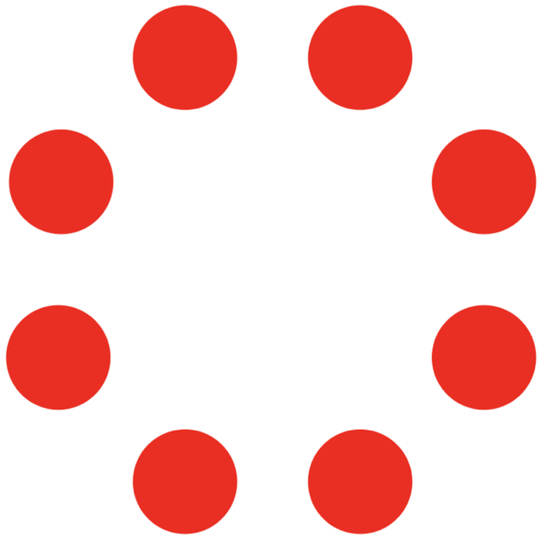
Watts and Strogatz - WS model

Number of nodes
in space

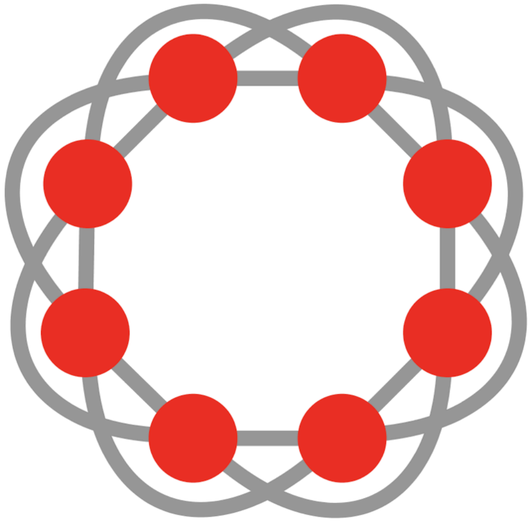


Watts and Strogatz - WS model

Number of nodes
in space

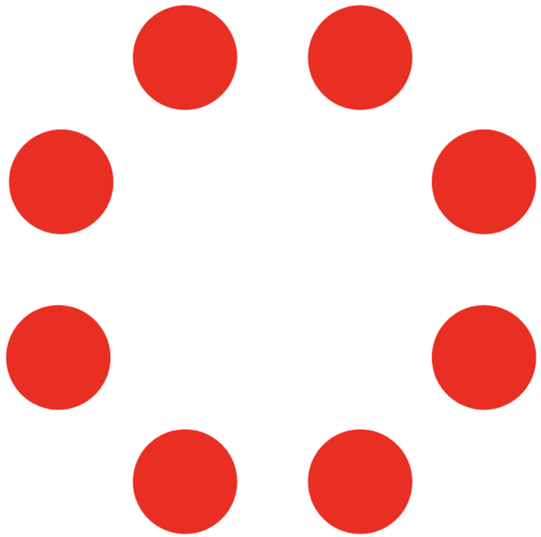


Connect nodes with
n nearest neighbours

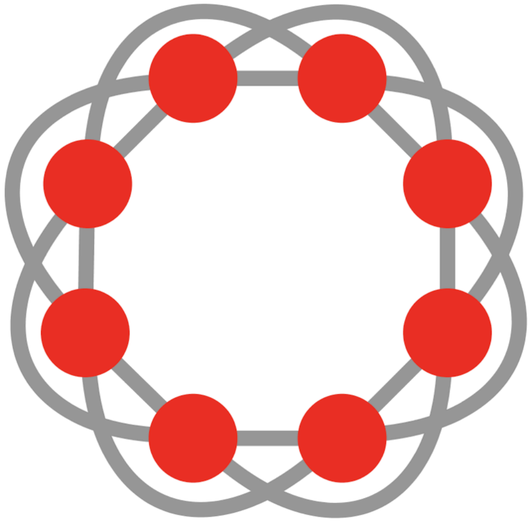


Watts and Strogatz - WS model

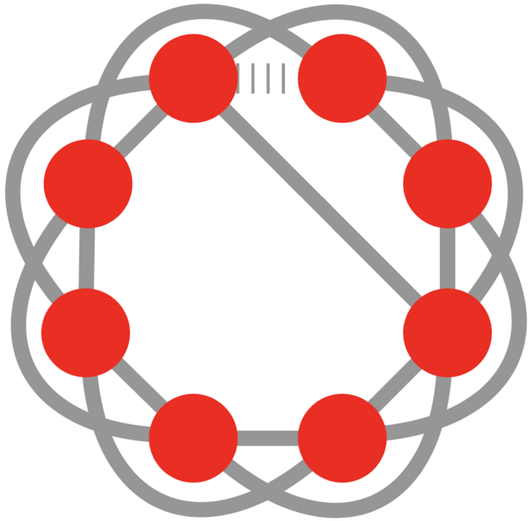
Number of nodes
in space



Connect nodes with
n nearest neighbours



Rewire links with probability *p*
(or add instead of rewiring)

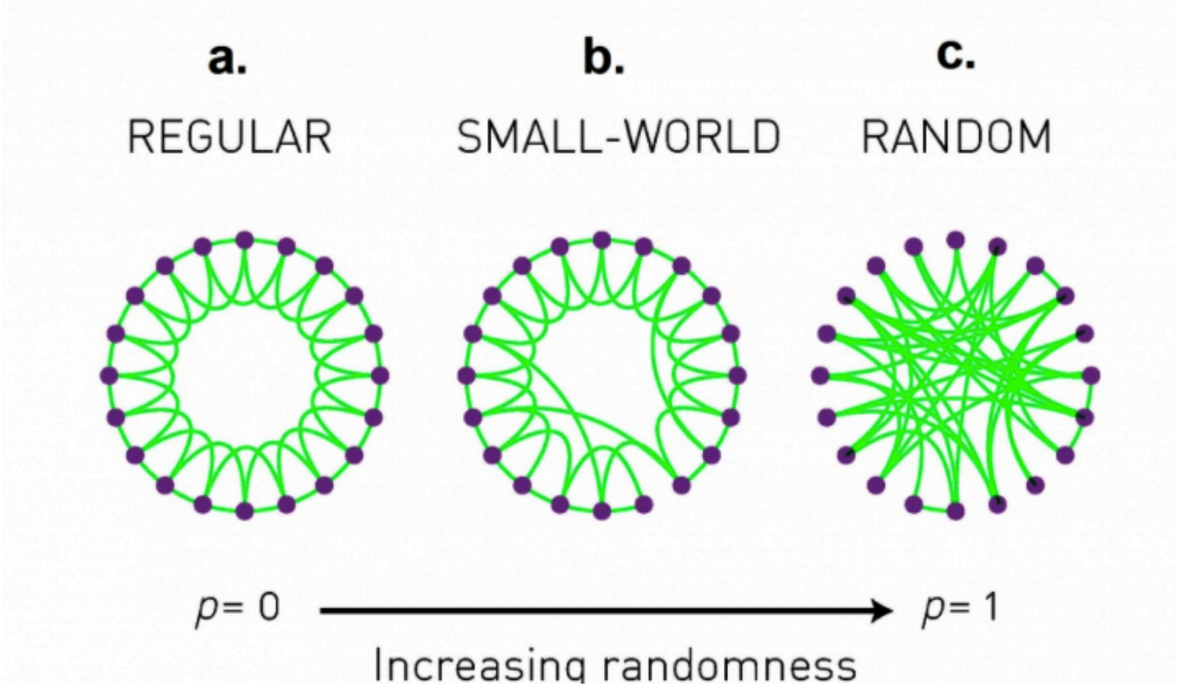


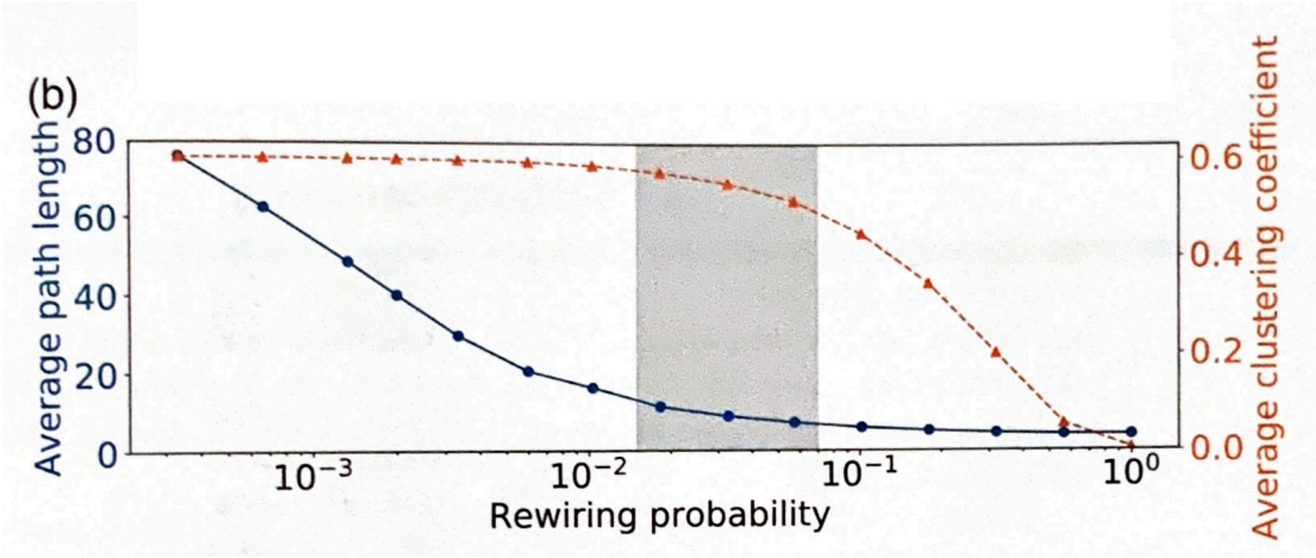
What do you get?

One component, Short paths, High clustering

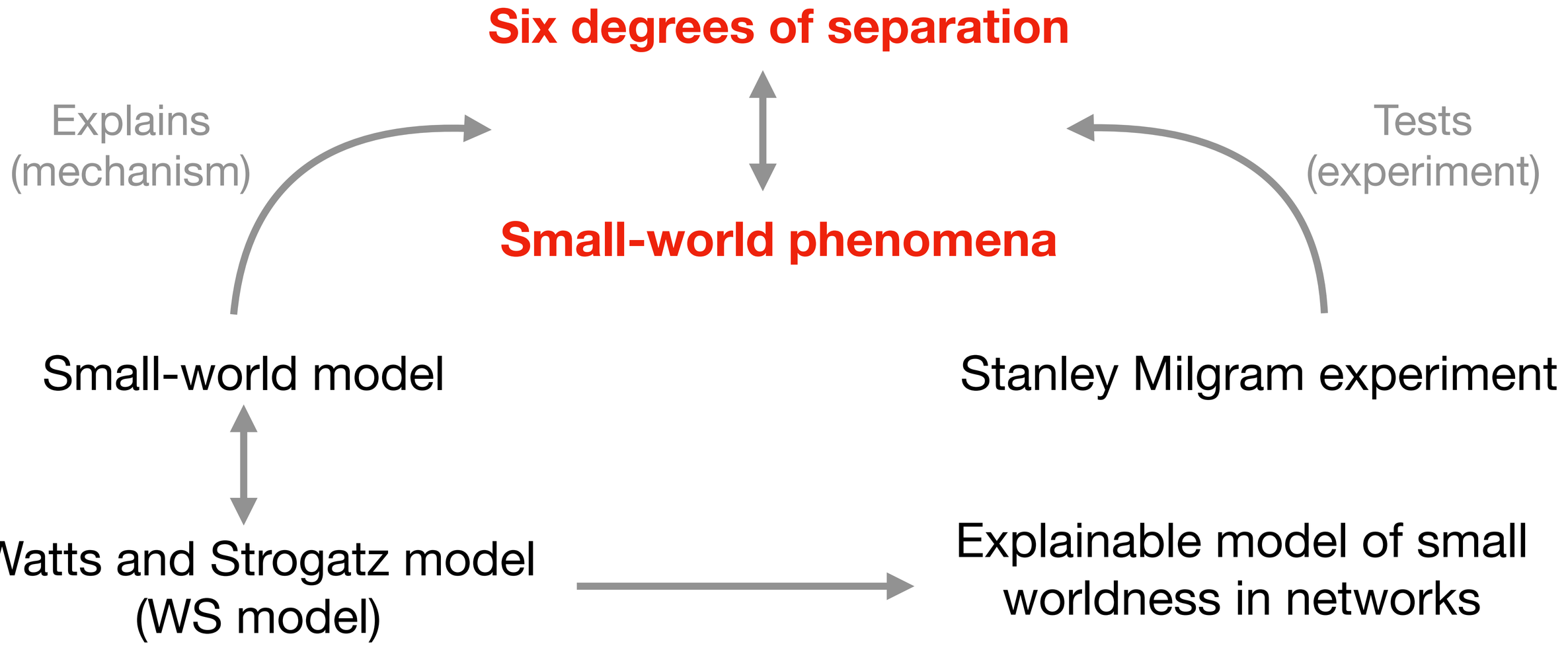
Watts and Strogatz - WS model

Why did Watts and Strogatz propose their model (WS-model)? In which way does it enhance the ER model?

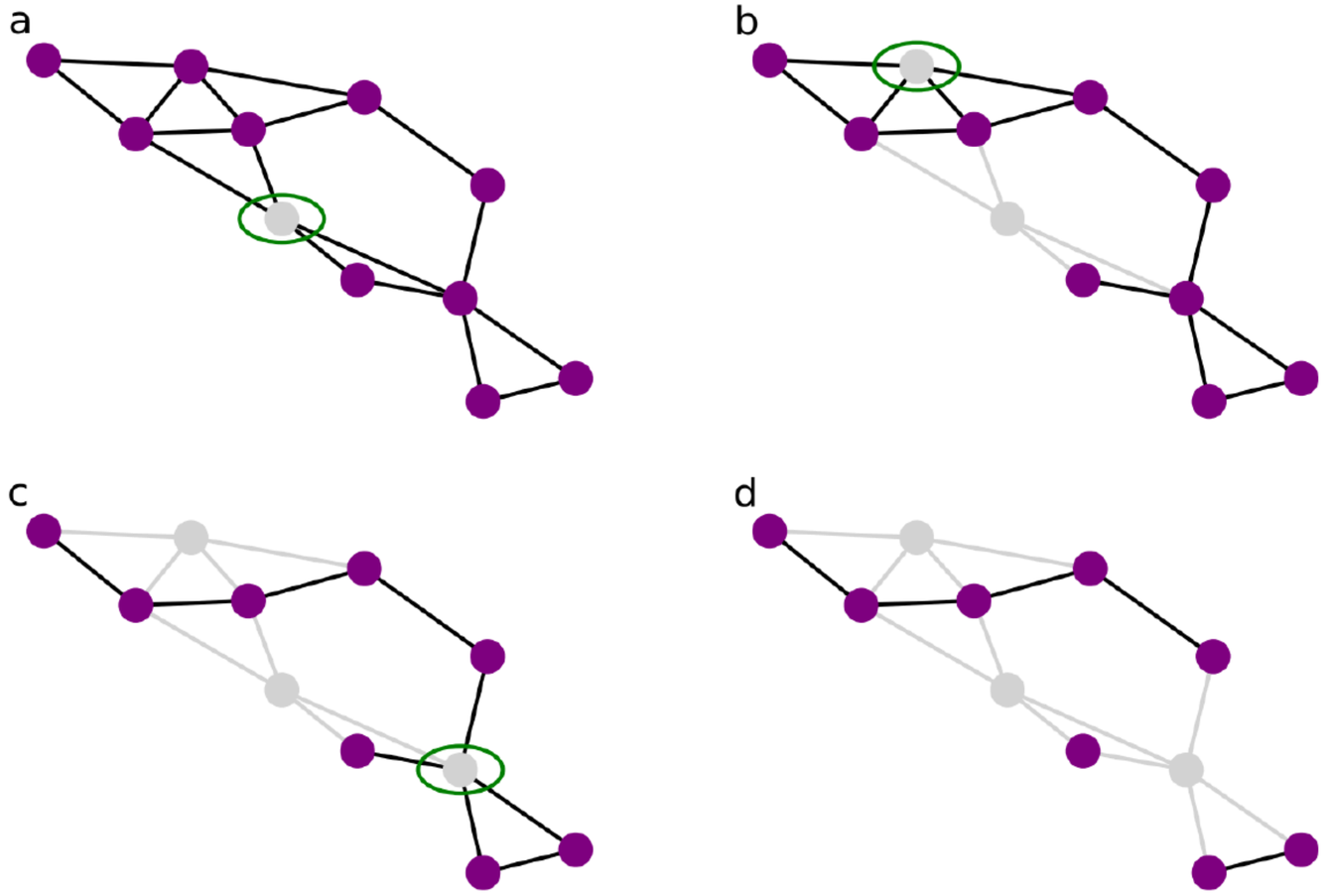




Disambiguation



Small world property - Robustness



What if we start removing nodes?

Small world property - Robustness

- Two strategies:

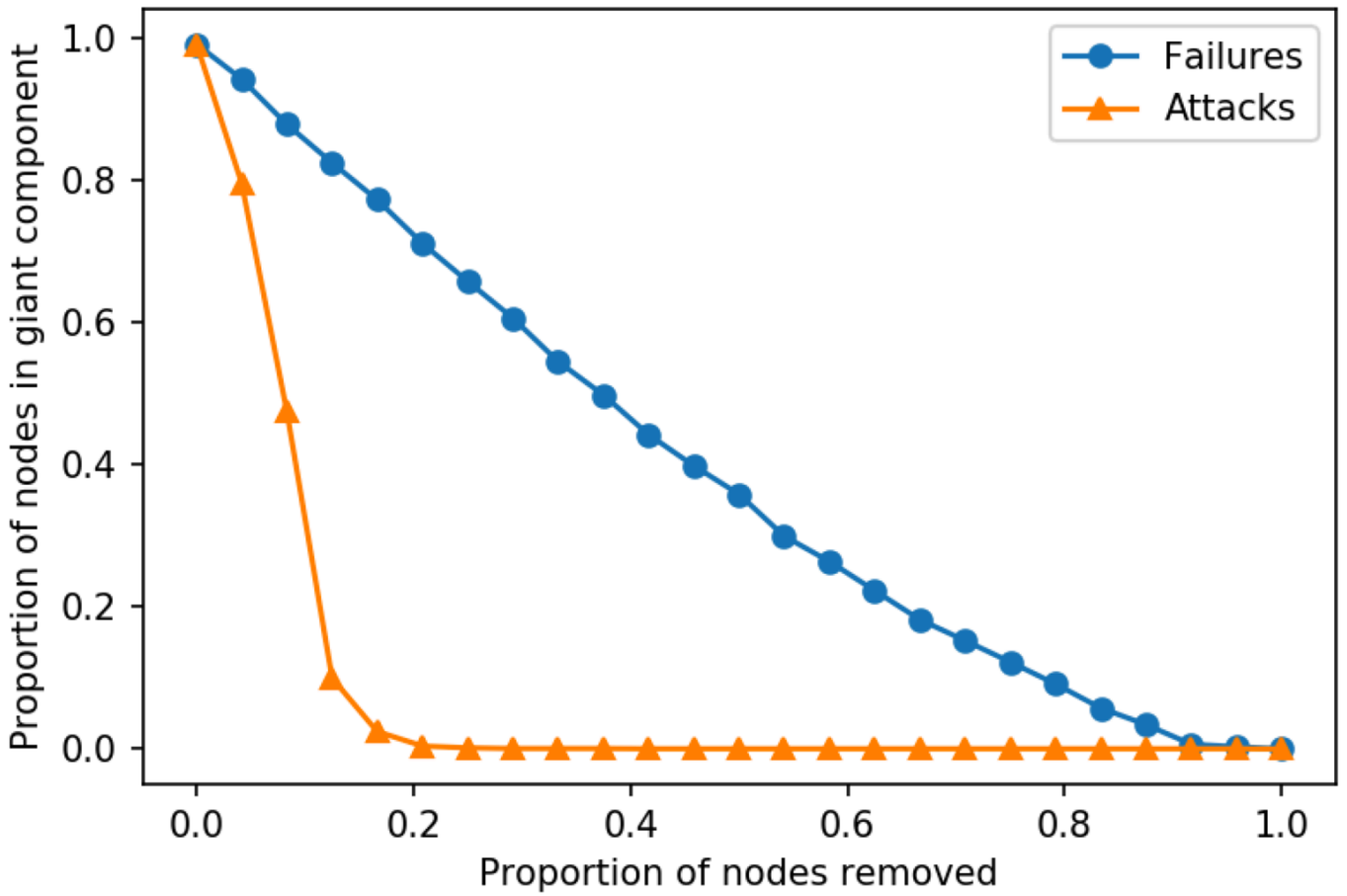
  1. **Random failures**: nodes break down randomly, so they are all chosen with the same probability

  2. **Attacks:** hubs are deliberately targeted — the larger the degree, the higher the probability of removing the node

- In the first approach, we remove a fraction *f* of nodes, chosen at random

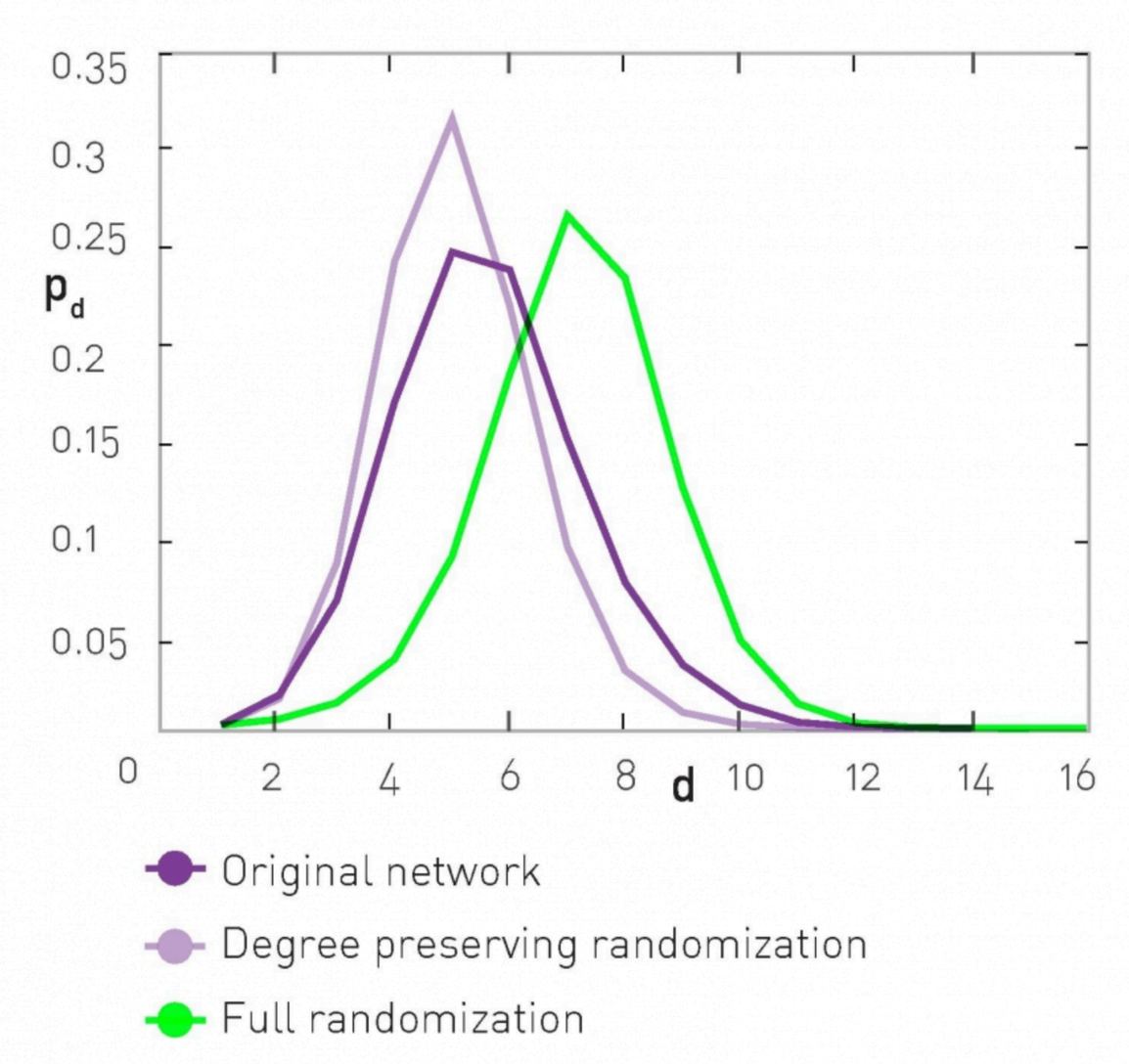
- In the second approach, we remove the fraction *f* of nodes with the largest degree, descending order.

Small world property - Robustness



Conclusion: real networks are robust against random failures but fragile against targeted attacks!

Network models, what for?



Explain mechanisms, behaviours, phenomena

We compare our real-world network behaviour to models (ER, SW, degree-preserving randomisation, etc.) as a **benchmark** for our results.

The model jungle

We saw only a couple of models; there are plenty more!

Model	Giant Component	Clustering	Degree distribution	…
Erdős Rényi	✓			
Small world	✓	✓		
Preferential attachement	✓		✓	
Newest fancy model				?

How to choose? Start with the **questions/hypotheses** you are making about your network and how to test them.


That's all Folks!