Universität Zürich[UZH]

Blockchain & Distributed Ledger Technologies

UZH Blockchain Center

Random Graph Models

Network Science - Lecture 5

Carlo Campajola

Blockchain & Distributed Ledger Technologies Group

Claudio J. Tessone

LAST WARNING:
from Assignment 3 (today) we will NOT accept non-standardised submissions

Lecture Objectives

1. Learn how to generate synthetic random networks
2. Get acquainted with the concept of network connectedness regimes
3. Learn the most important network randomisation algorithms

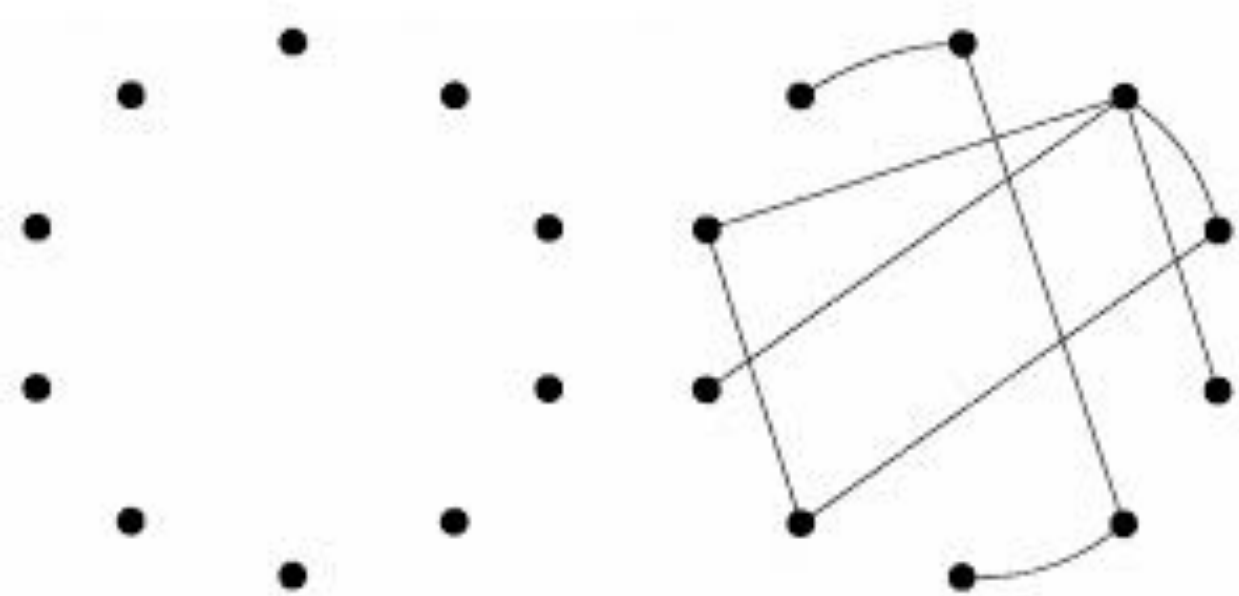
What is a random network?

A network resulting from a stochastic process

Example

A simple stochastic rule:

- N nodes;
- randomly select L pairs of nodes (without replacement)
- add a link on each of the pairs



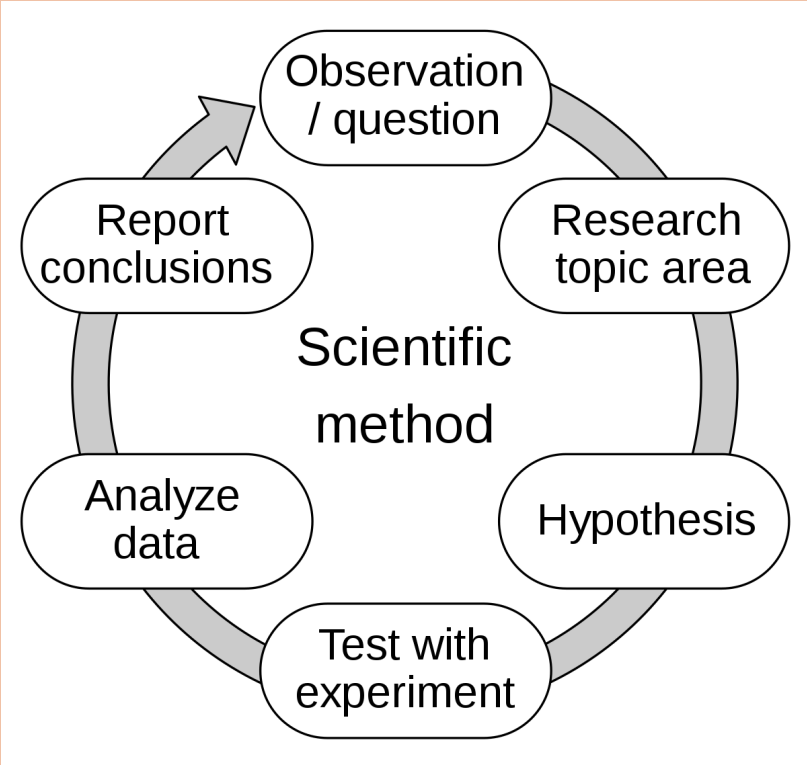
Why random networks?

Motivations

- Statistical testing (null models): how likely is an observation if the network is formed with some rule?
- Network generation: sometimes you just need to decide how to connect some nodes - e.g. p2p networks, agent-based models, ...

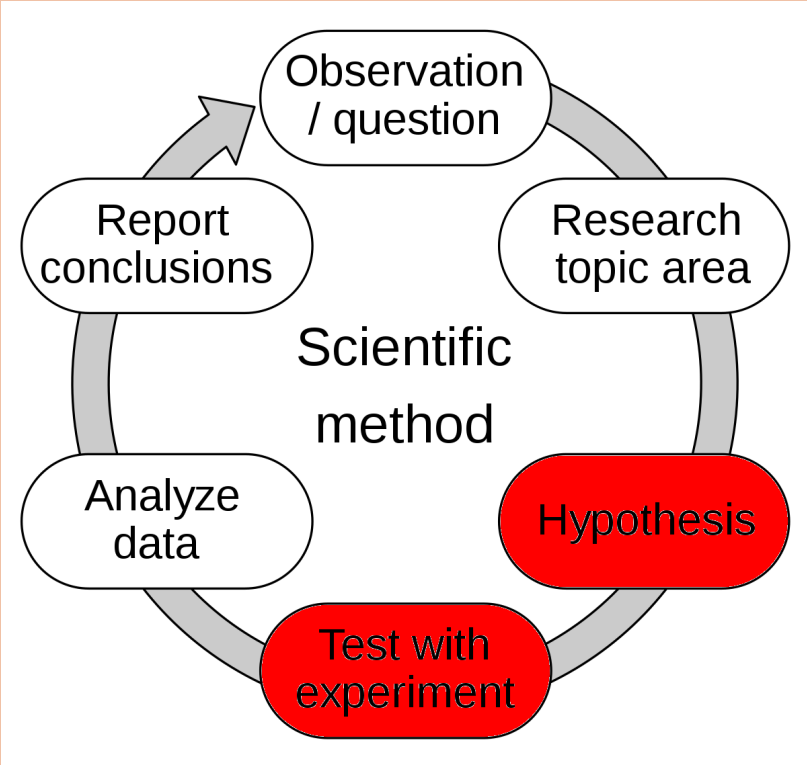
Refresher: Hypothesis testing

The scientific method is based on **falsification**



Refresher: Hypothesis testing

The scientific method is based on **falsification**



Refresher: Hypothesis testing

The scientific method is based on **falsification**

1. Make a simple null hypothesis
2. What are the consequences of your hypothesis?
3. Identify a probability distribution for observations if your null hypothesis was true
4. Calculate the probability of observing something as extreme as the data, given your hypothesis
5. If the probability is low, the hypothesis is rejected and an alternative should be considered

Refresher: Hypothesis testing

The scientific method is based on **falsification**

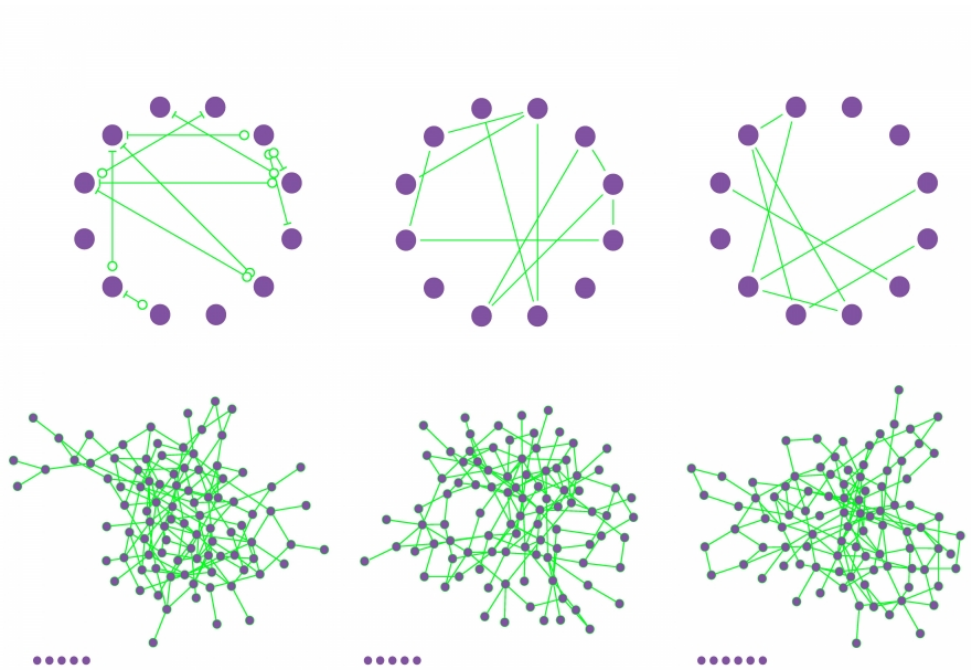
1. Null hypothesis $\mathbb{H}_0$: nodes are linked by random selection of pairs
2. Consequence: the maximum degree follows from the rule above
3. Find the null max degree probability $\mathbb{P}(k_{max}|\mathbb{H}_0)$
4. Measure $\hat{k}_{max}$ on data and calculate $p = \mathbb{P}(k_{max} \geq \hat{k}_{max}|\mathbb{H}_0)$
5. If $p < \alpha$ then reject $\mathbb{H}_0$ at confidence level α: random linkage is not realistic

The Erdos-Renyi Model

- First attempt to approximate a real network (null model in multiple scenarios)
- **Network density is fixed**
- Connections are placed just randomly between every pair of nodes

  - $G(N, L)$ *Model:* N *labeled nodes are connected with* L *randomly placed links*
  - $G(N, p)$ *Model: each pair of* N *labeled nodes is connected with* ***edge creation probability*** p

Erdos-Renyi is "homogeneous"

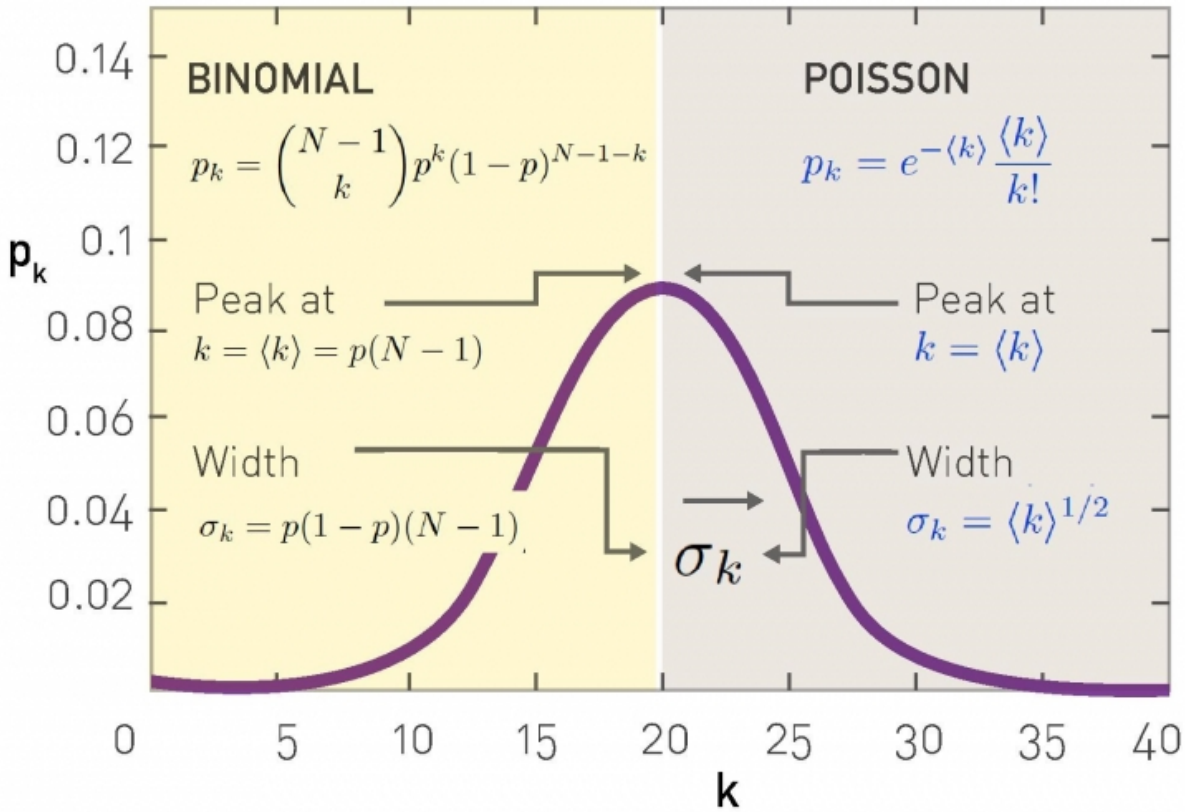


- Links are equally likely between pairs of nodes
- There are no nodes more important than others

Degree distribution

- The binomial distribution describes the number of successes in N independent experiments with two possible outcomes, in which the probability of one outcome is p, and of the other is $1-p$.
- Exact distribution: Binomial: $P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$
- Depends on the network size
- Approximation: Poisson: $P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$
- Does not depend on the network size, only on the average degree
- Approximation is valid when $\langle k \rangle \ll N$

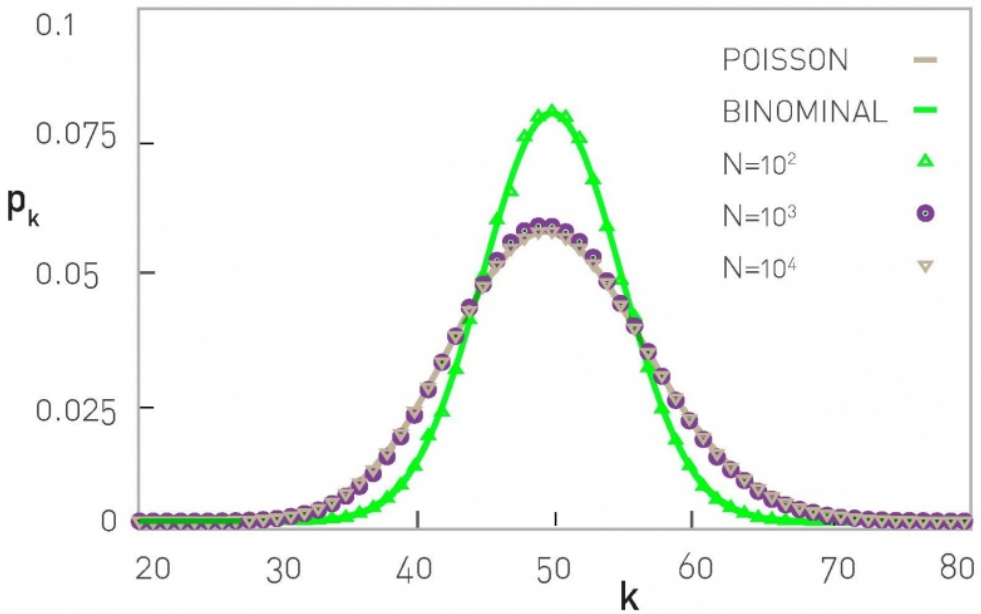
Binomial vs. Poisson



Poisson is an approximation

As network size N goes to ∞, for fixed $\langle k \rangle = pN$:

$$\binom{N-1}{k} p^k (1-p)^{N-1-k} \longrightarrow e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$



Giant component

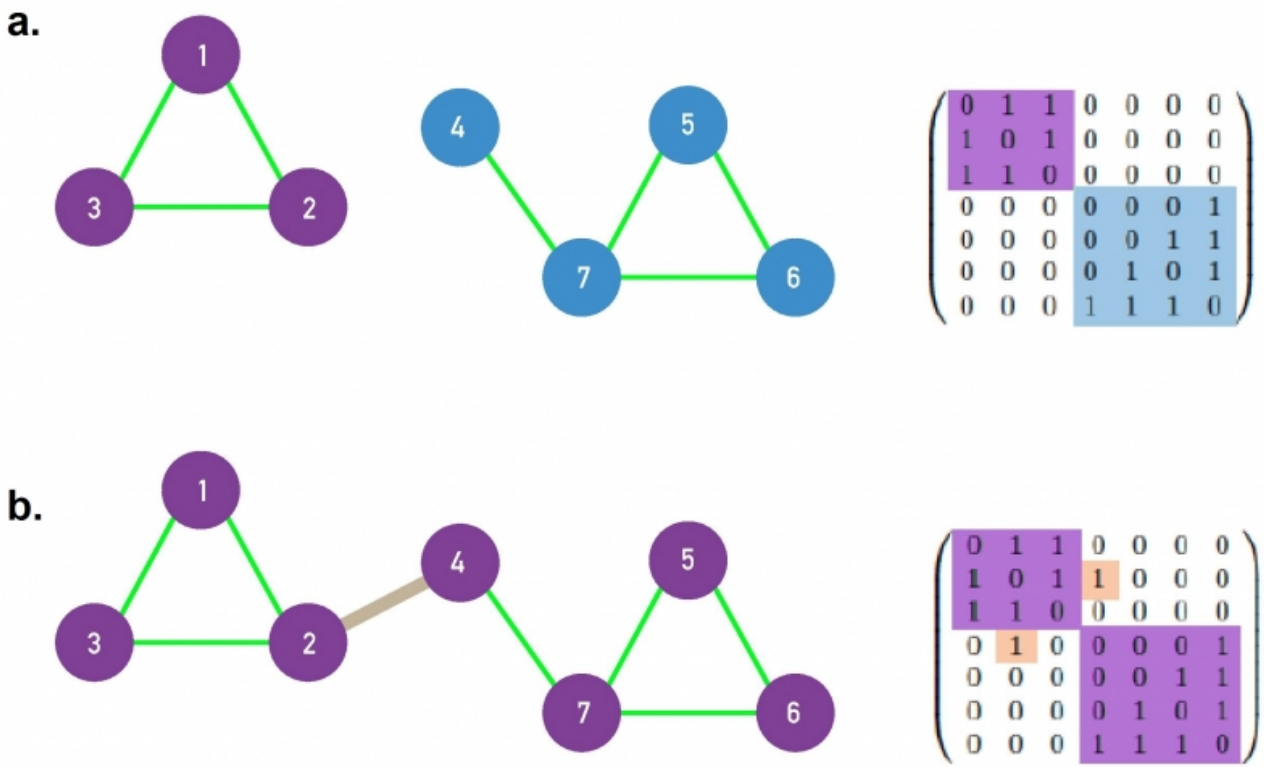
Connectedness

- The network behind the phone or the Internet must be capable of establishing a path between any two nodes
- Key utility of most networks: they ensure connectedness
- In an undirected network nodes i and j are connected if there is (at least) a path between them
- They are disconnected if such a path does not exist, in which case we have $d_{ij} = \infty$

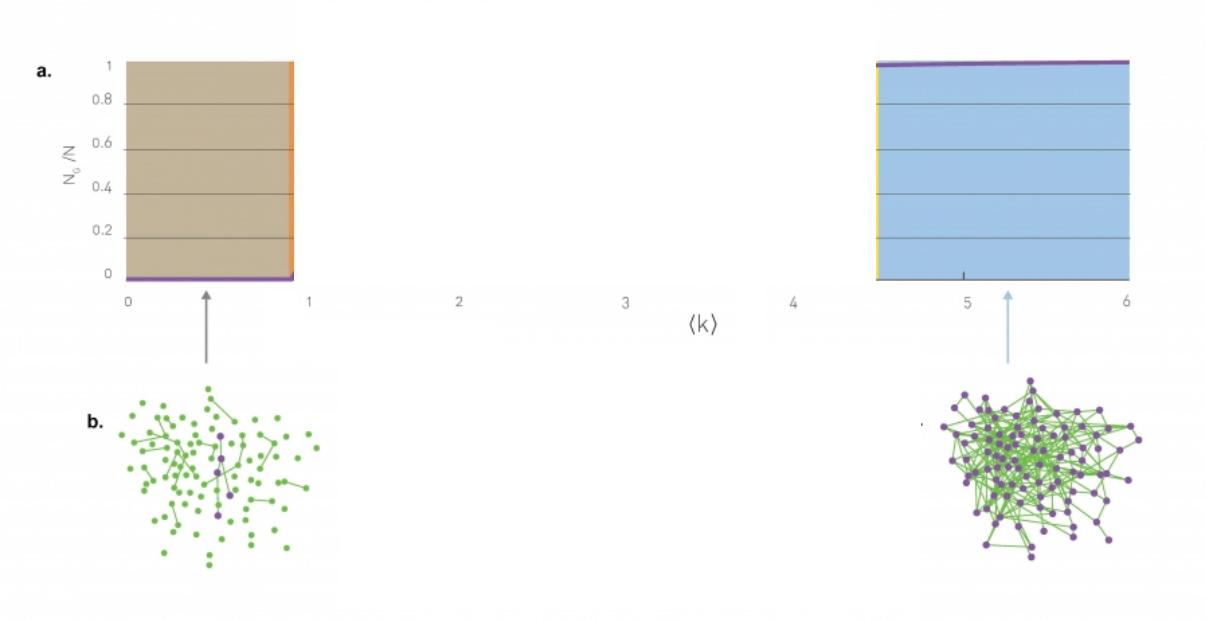
Connectedness

- A network is **connected** if all pairs of nodes in the network are connected
- A network is **disconnected** if there is at least one pair with $d_{ij} = \infty$
- A **component** is a subset of nodes in a network, so that there is a path between any two nodes that belong to the component, but one cannot add any more nodes to it that would have the same property.

Connectedness



Evolution of random networks



- N_G = size of the largest connected component
- If $p \sim 0 \Rightarrow \langle k \rangle \sim 0$. $N_G \sim 1$, $\frac{N_G}{N} \to 0$ for large N;
- If $p \sim 1 \Rightarrow \langle k \rangle \sim N$. $N_G \sim N$, $\frac{N_G}{N} \to 1$ for large N;
- What happens in between?

Giant Component

- We call a **giant component** a network component whose size grows *proportionally* with N
- The size of giant component does not grow linearly with p
- Instead, there is a *phase transition* from constant size to extensive size
- For a Poisson random graph, the size of the giant component is defined by the equation:

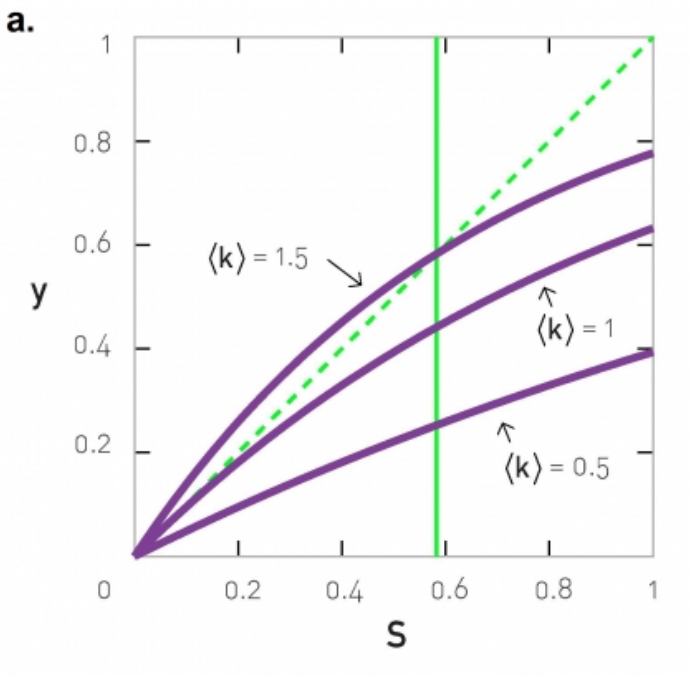
Size of the giant component

Equation given by Erdos and Rényi in 1959:

$$N_G = 1 - e^{-\langle k \rangle N_G}$$

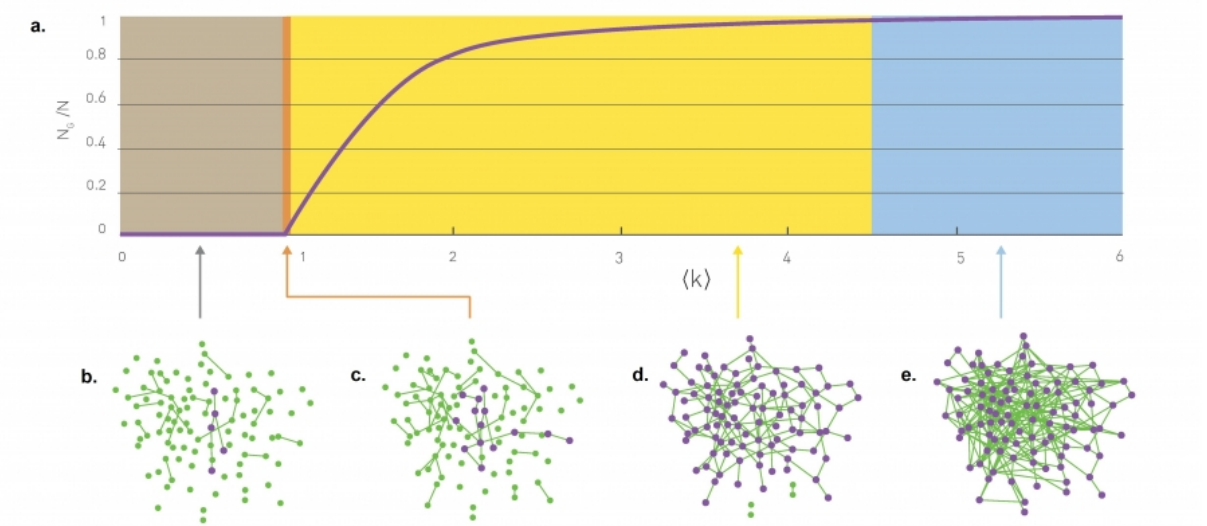
Erdos-Rényi equation

- There is no analytical solution to this equation
- We can take a look at the graphical solution for $S = N_G/N$



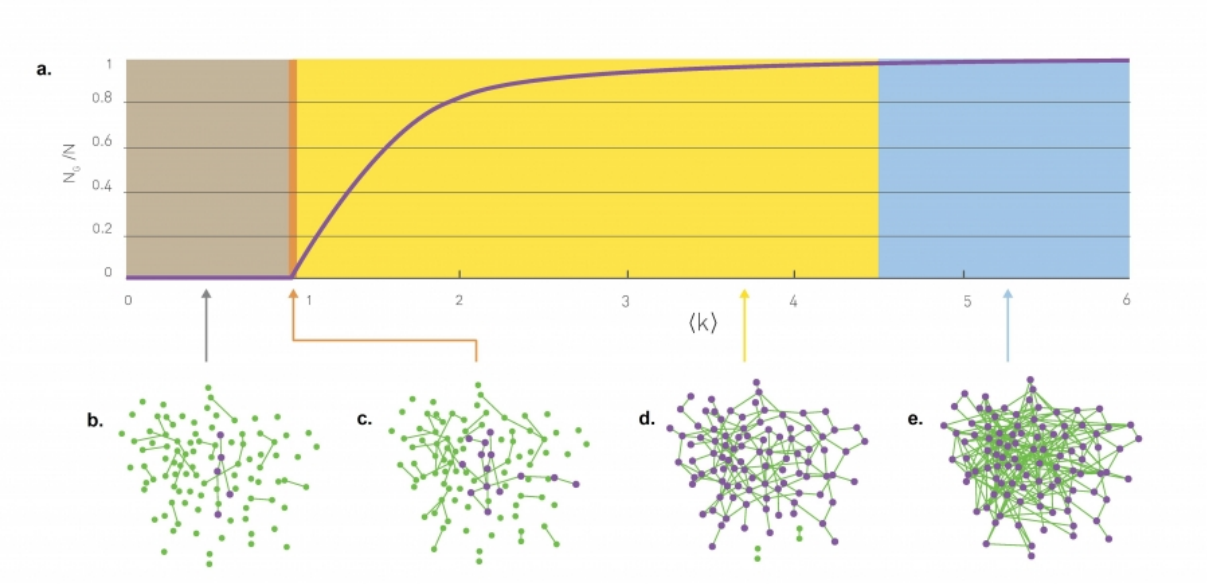
- Depending on value of $\langle k \rangle$, there can be one or two solutions.
- $\langle k \rangle = 1$ is the threshold between the regimes with and without a non-trivial solution.

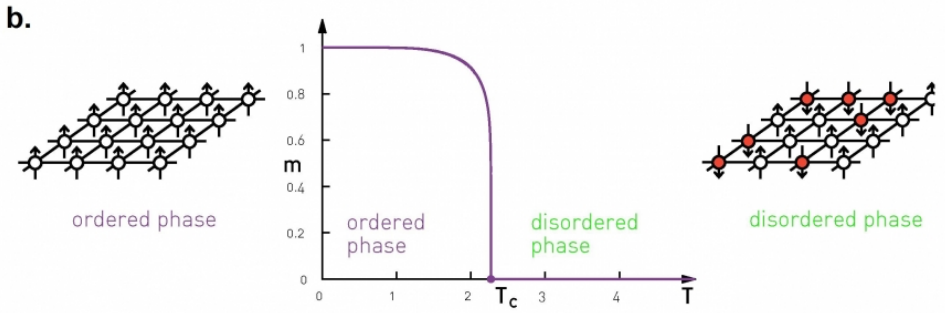
Giant Component



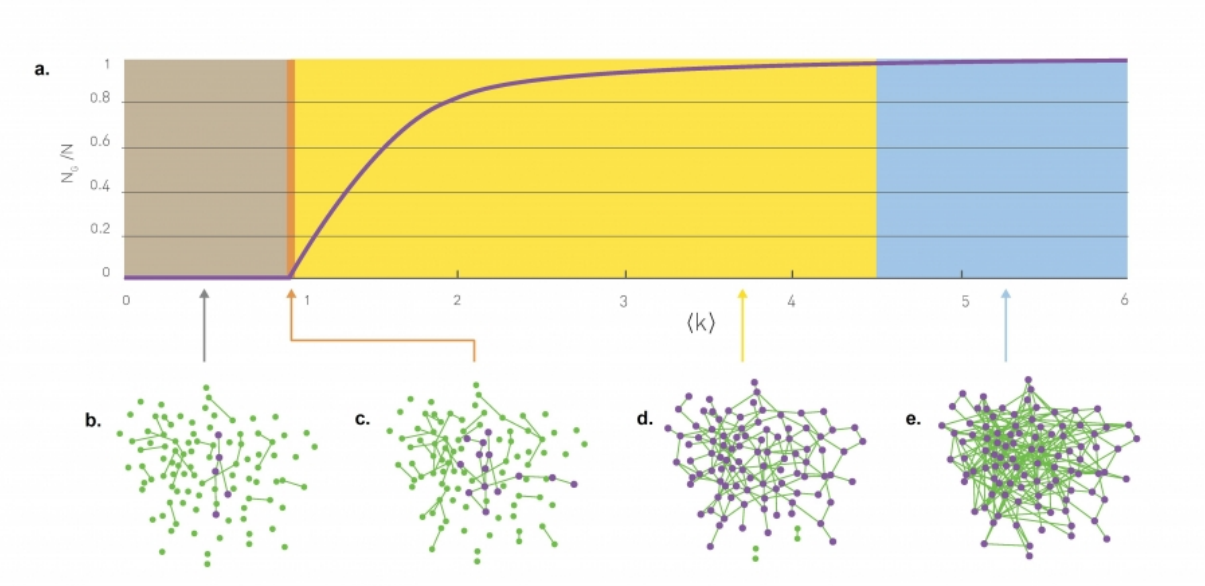
- $\frac{N_G}{N}$ as function of average degree $\langle k \rangle$.
- Phase transition at $\langle k \rangle = 1$, emergence of a giant component.
- Sample network and its properties in the four regimes that characterise a random network.

Network regimes



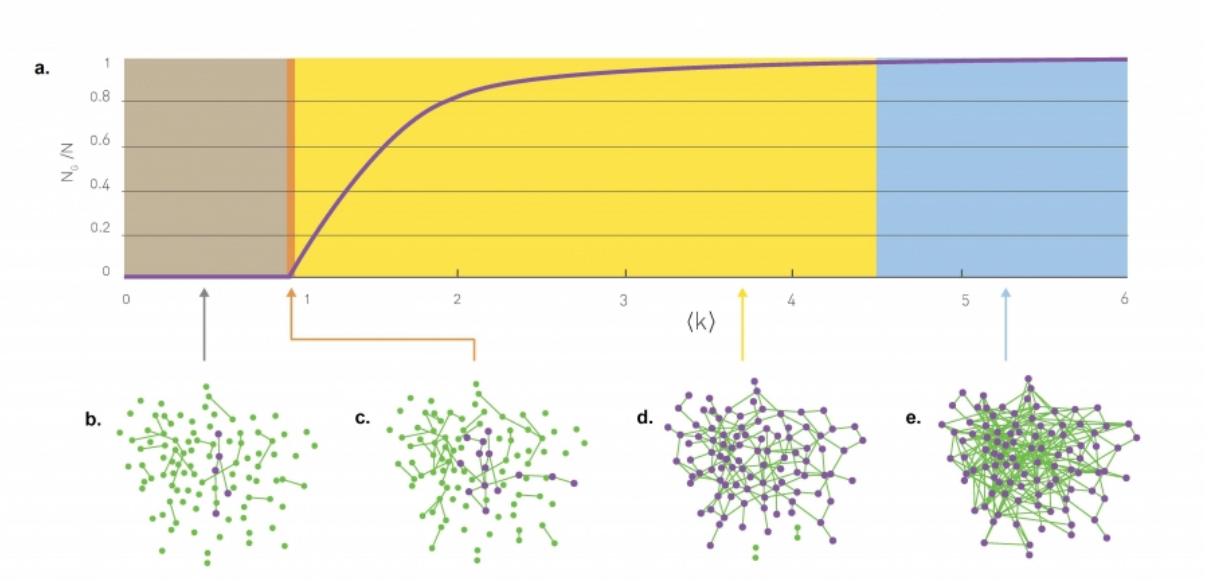


Network regimes



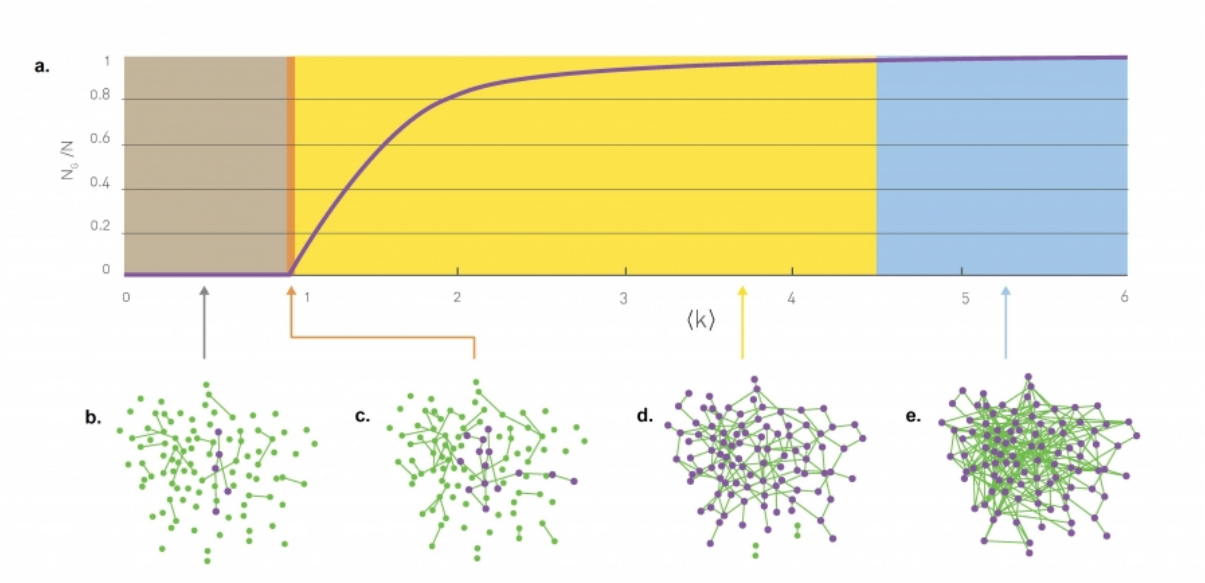
+ **Subcritical Regime**: $\langle k \rangle < 1 \iff p < \frac{1}{N}$:
  numerous tiny components, fig. b

Network regimes



+ **Critical point**: $\langle k \rangle = 1 \iff p = \frac{1}{N} = p_{crit}$:
  Still no giant component, the size of the largest component is $N_G \propto N^{\frac{2}{3}}$, fig. c

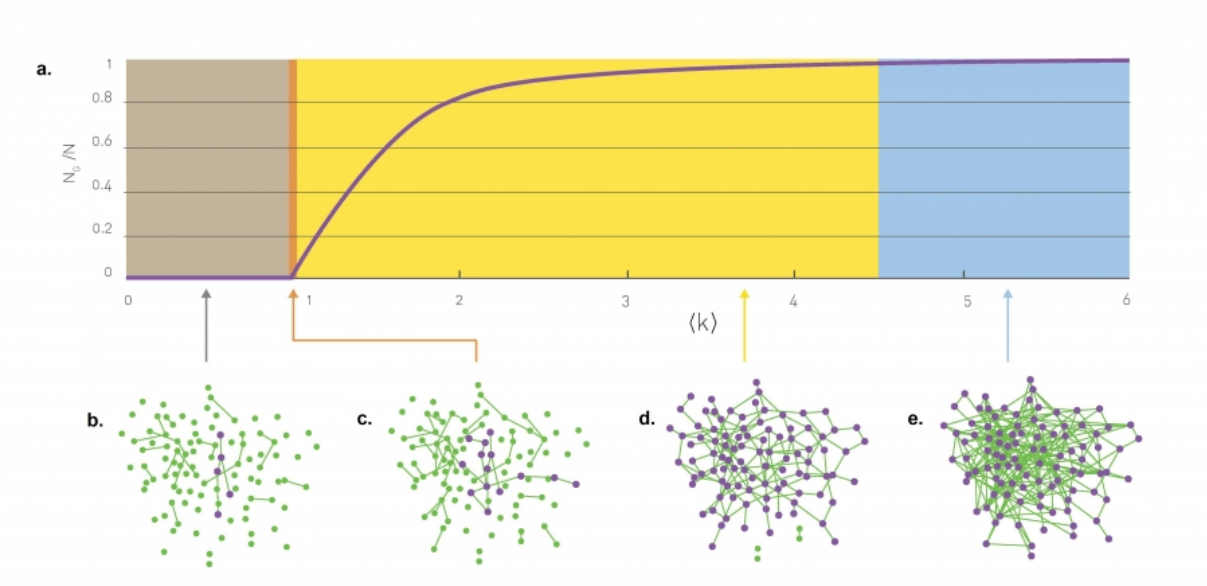
Network regimes



+ **Supercritical Regime**: $\langle k \rangle > 1 \iff p > \frac{1}{N}$

  Giant component size is $N_G \propto (p - p_{crit})N$, fig. d

Network regimes



+ **Connected Regime**: $\langle k \rangle > \log(N)$

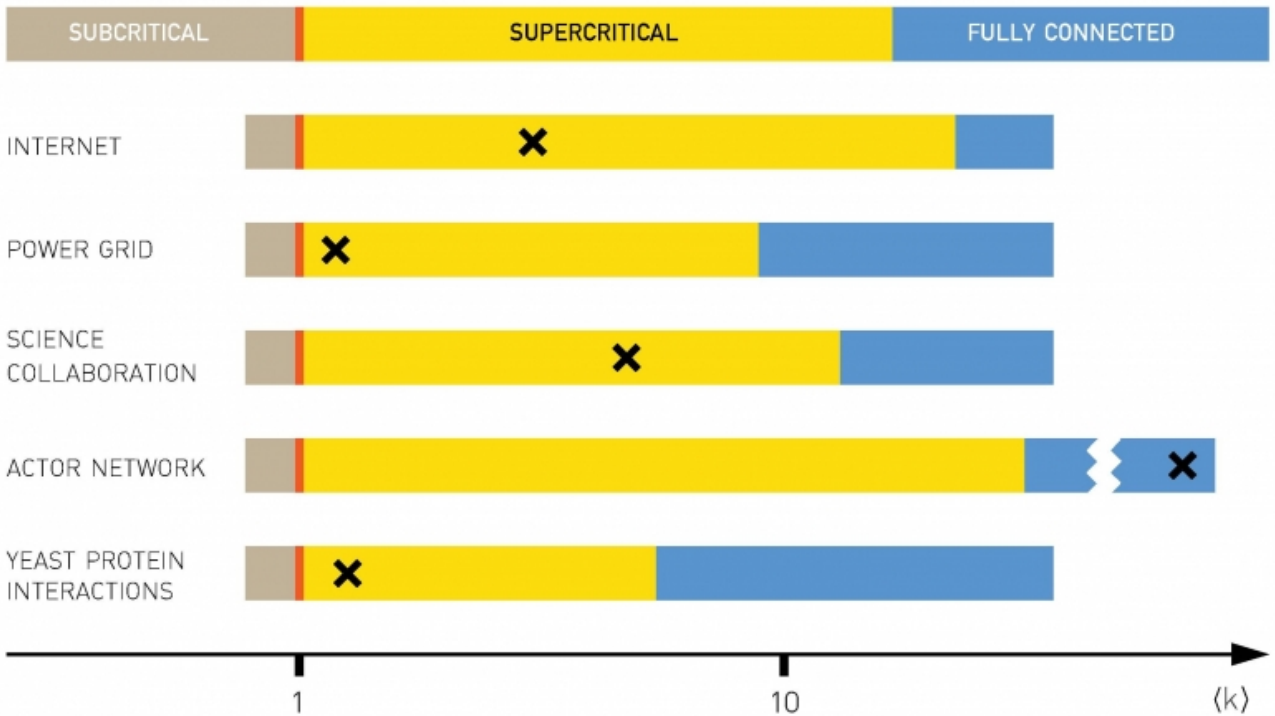
  All components are absorbed by the giant component, fig. e

Network regimes

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

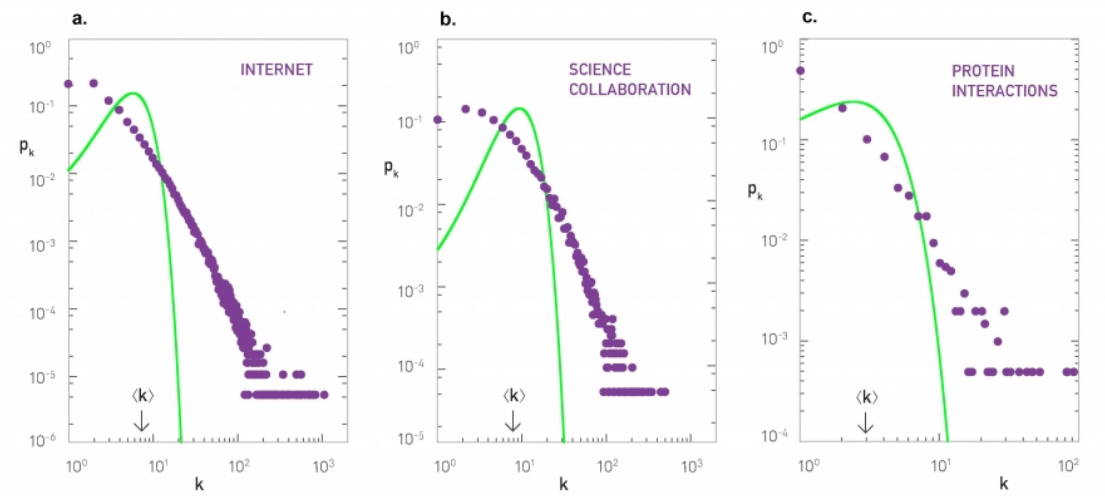
Real networks regimes: predictions

Note: Only valid if the networks are described by ER model.



Real networks are not Poisson

Figure: The green line corresponds to the Poisson prediction, obtained by measuring $\langle k \rangle$ for the real network and then plotting Poisson curve.



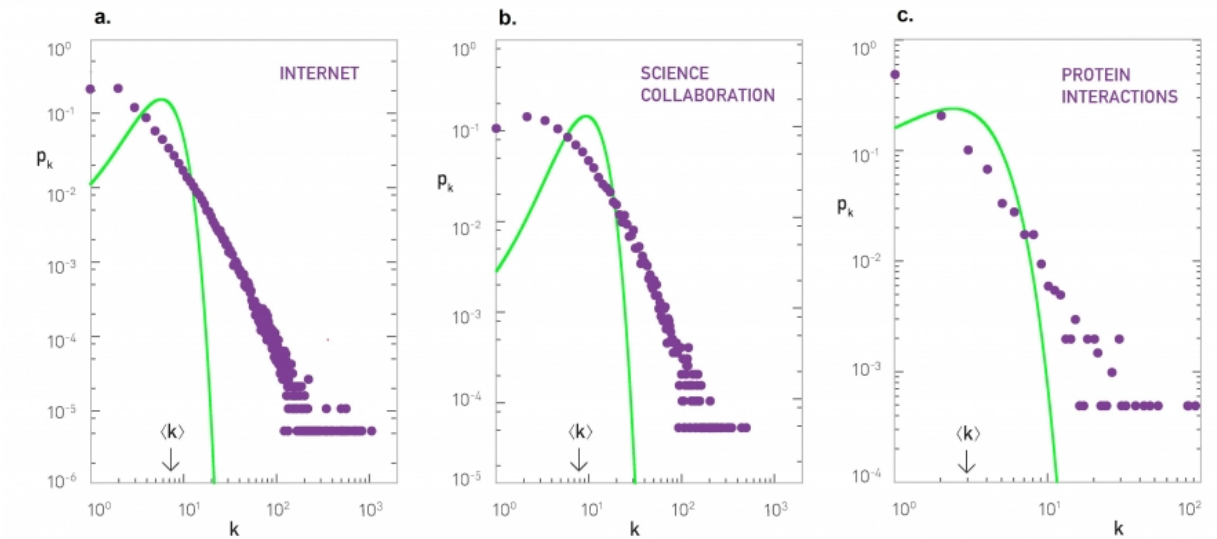
Random network model:

- *underestimates* size and frequency of high degree nodes;
- *underestimates* the number of low degree nodes.
- *overestimates* number of nodes in the vicinity of $\langle k \rangle$.

Configuration Model

ER versus configuration model

- Erdos-Rényi model only replicated network density
- What happens if we want to generate networks with the same degree distribution than one observation?
- Final question is: How much of the network properties are a result of the degree distribution?

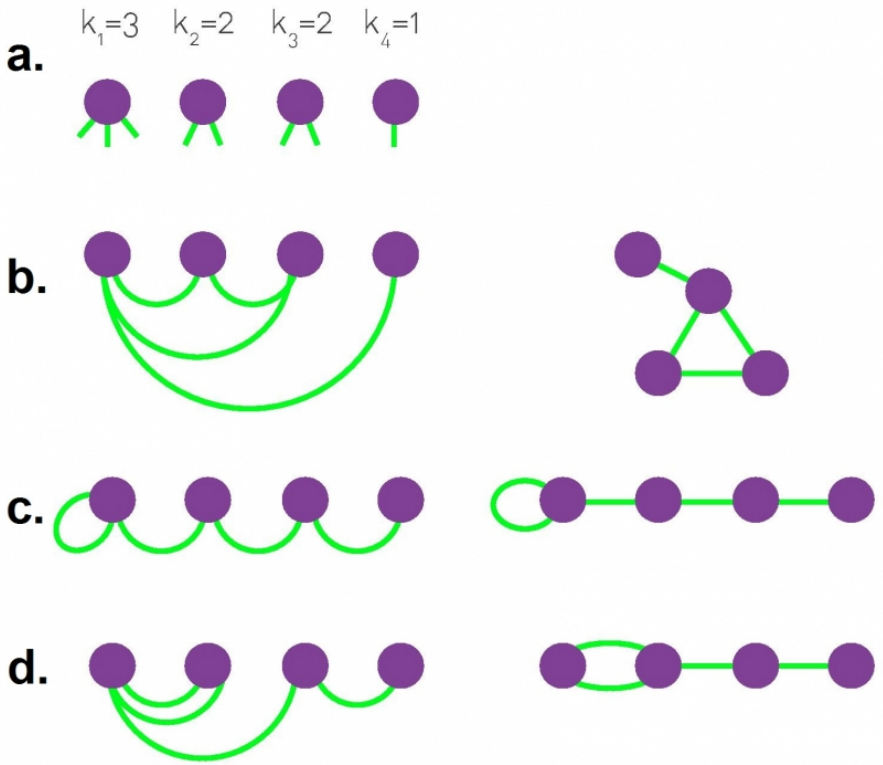


Configuration model algorithm

Algorithm:

1. Generate the degree sequence either analytically from a preselected $P(k)$ distribution or extracted from a real network
2. Assign a degree to each node, represented as *stubs* or half-links
3. Start from an even number of stubs
4. Randomly select a stub pair and connect them
5. Then randomly choose another pair from the remaining $2L-2$ stubs and connect them
6. This procedure is repeated until all stubs are paired up
7. Depending on the order in which the stubs were chosen, different networks are obtained

Configuration model illustration



Configuration model properties

- Helps to build a network with a pre-defined degree **sequence**
- This **fixes** the number of **edges** in the network (like ER models)
- This is an extension to the $G(N, L)$ model, which also fixes the number of edges L
- Generates each possible matching of stubs with **equal probability**
- Each stub is equally likely to be connected to any other.
- All generated networks are **equally likely** (with self-loops or multiedges)
- *Requirement*: start with an even total number of stubs
- *Note*: self-loops and multiedges cannot be avoided

Configuration model: edge probability

- Assume $k_i, k_j > 0$
- L edges $\rightarrow 2L$ stubs
- Any stub out of k_i stubs of node i can lead to any $2L-1$ remaining ones
- There are k_j possibilities to connect to node j.
- Thus: $p_{ij} = \frac{k_i k_j}{2L-1}$
- Also correct for nodes with no stubs.

Hidden Parameter Model

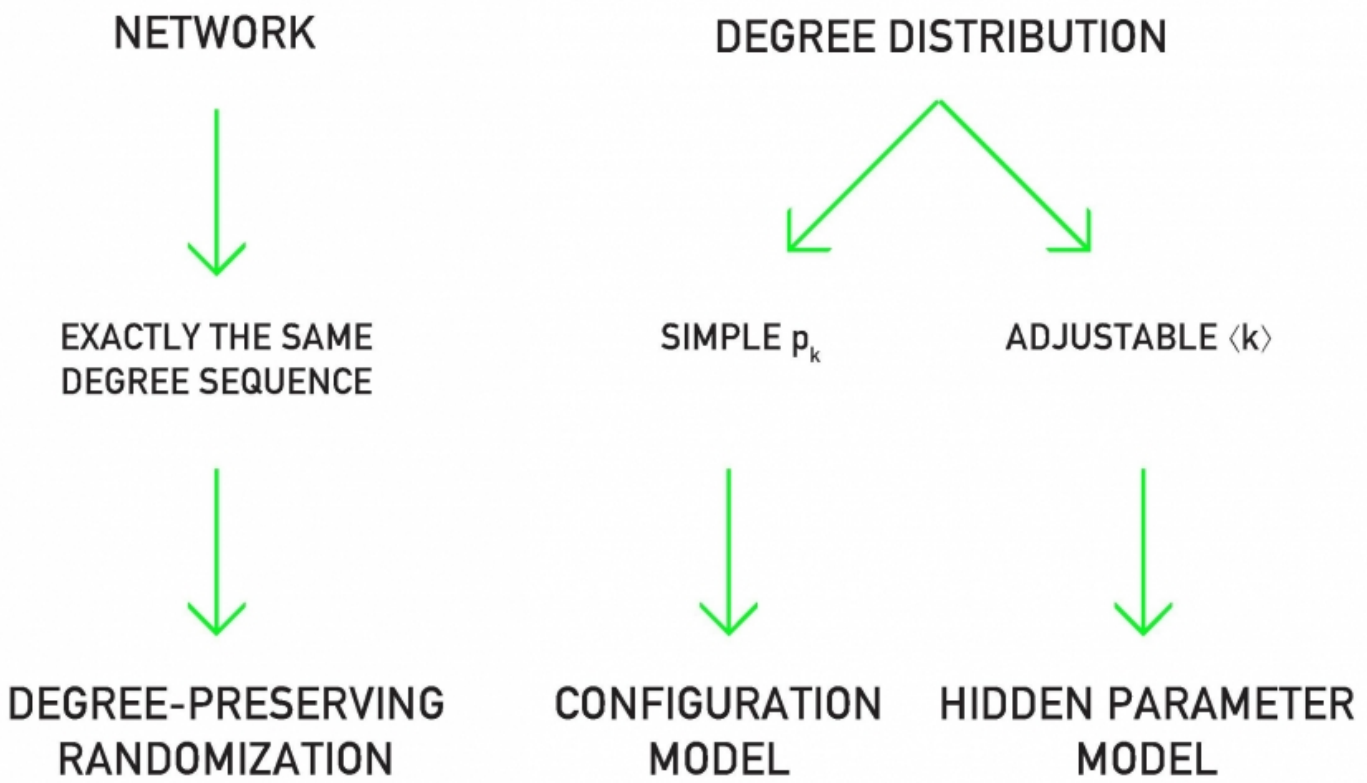
- Generalisation to $G(N,p)$ model, which fixes probability of an edge.
- We define a *hidden parameter* $\eta_i \sim \rho(\eta)$ for each vertex.
- Each node pair is connected with probability $p(\eta_i, \eta_j) = \frac{\eta_i \eta_j}{\langle \eta \rangle N}$
- The degree distribution of the obtained network is:
  - Analytical distribution $\rho(\eta)$:
    $p_k = \int \frac{e^{-\eta}\eta^k}{k!}\rho(\eta)d\eta$
  - Determenistic sequence $\eta_i \in \{\eta_1, \eta_2, ..., \eta_N\}$:
    $p_k = \frac{1}{N}\sum_j \frac{e^{-\eta_j}\eta_j^k}{k!}$
- Special case: choosing $\eta_i = \frac{c}{i^\alpha}$ leads to *scale-free* network with $p_k \sim k^{-(1+\frac{1}{\alpha})}$

Degree-preserving randomisation

Degree-preserving randomisation

- Used when we check if a certain network property is *predicted by its degree distribution alone*, or if it represents some additional property not contained in $P(k)$.
- Generates networks that are wired randomly, but whose $P(k)$ is **identical** to the original network.
- **Idea**: we randomly select two links and swap them, if this does not lead to multilink.
- The **degree** of each of the **four** involved nodes remains **unchanged**.
- Different from *full randomisation*, where links are swapped without preserving the degree.

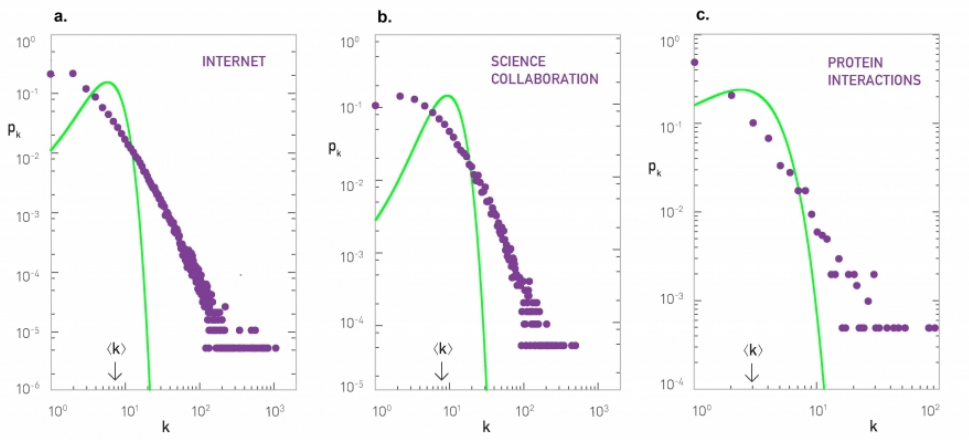
Choosing the generation algorithm



Small World

Real networks are not Poisson

Figure: The green line corresponds to the Poisson prediction, obtained by measuring $\langle k \rangle$ for the real network and then plotting Poisson curve.



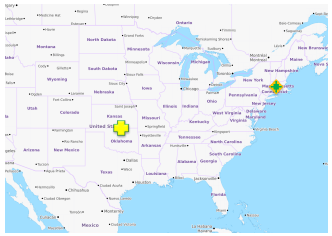
Random network model:

- *underestimates* size and frequency of high degree nodes;
- *underestimates* the number of low degree nodes.
- *overestimates* number of nodes in the vicinity of $\langle k \rangle$.

Theory of 6 handshakes

- Pool and Kochen, 1950:
  What is the probability that two strangers will have a mutual friend? What about when there is no mutual friend – how long would the chain of intermediaries be?
- Milgram, 1969:
  Let's conduct an experiment!
- John Guare, 1990:
  There are six degree of separation between you and everyone else on this planet
- Backstrom et al., 2012:
  Four Degrees of Separation

Milgram's experiment



- 300 initial senders in the central states of the USA: *Letters and business reply cards*
- The letters should reach a specific person in Boston
- *If a sender does not know how to reach this person, she should forward the letter to some acquaintance she thought may know how to deliver it*
- Same instructions apply to the next person

Milgram's experiment: results

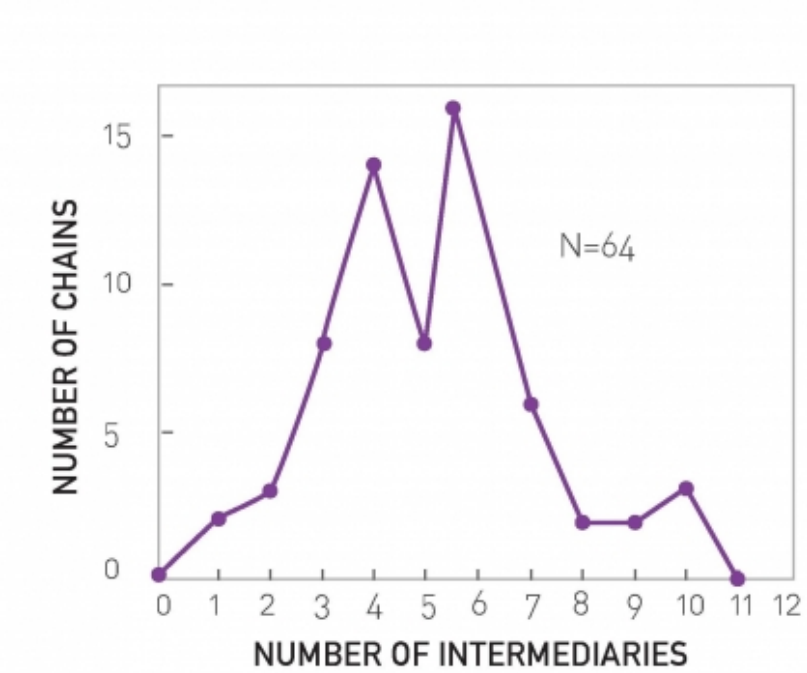
- Around $\frac{1}{5}$ of all letters reached the target
- **Among the arrived letters, the average number of intermediate steps was between 5 and 6 (over many trials)**
- This research suggested that human society is characterised by short path lengths: we are all linked by short chains of acquaintances, or "six degrees of separation".
- The small-world phenomenon is a fundamental property of social networks

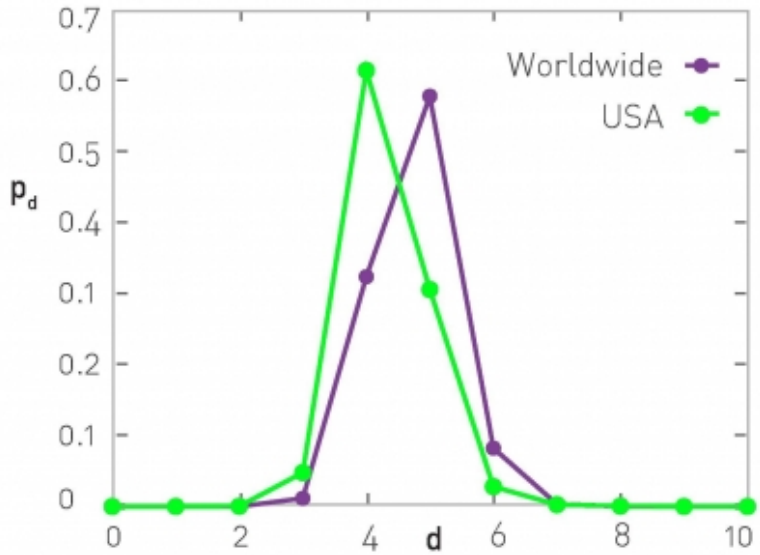
Milgram's experiment: results

- Around $\frac{1}{5}$ of all letters reached the target
- **Among the arrived letters, the average number of intermediate steps was between 5 and 6 (over many trials)**
- This research suggested that human society is characterised by short path lengths: we are all linked by short chains of acquaintances, or "six degrees of separation".
- The small-world phenomenon is a fundamental property of social networks

Milgram's experiment: results

[Albert- László Barabási, *Network Science* (Cambridge University Press, 2015)]





Small-world phenomenon

Imagine a network where a node has on average $\langle k \rangle$ connections.
Then:

- There are $\langle k \rangle$ nodes at distance 1; 10^3 acquaintences
- $\langle k \rangle^2$ nodes at distance 2; 10^6 acquaintences
- $\langle k \rangle^3$ nodes at distance 3. 10^9 acquaintences

  ...
- $\langle k \rangle^d$ nodes at distance d. 10^{3d} acquaintences

For the human acquaintences network, the estimate for $\langle k \rangle$ is approximately 1000.
This means that we have 10^6 acquaintences at distance 2, about a billion at distance 3

Small-world phenomenon

Expected number of nodes up to distance d from a starting node:

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

$N(d)$ must not exceed the total size of the network N:

$$\langle k \rangle^d \approx N(d) \leq N \tag{2}$$

which gives:

$$d \leq \frac{\log N}{\log \langle k \rangle} \tag{3}$$

The problem of this argument is that it does not account for clustering. (tree-like approximation)

Small-world phenomenon

The **small-world** phenomenon implies that *the distance between two randomly chosen nodes in a network is short*.

Short compared to what?

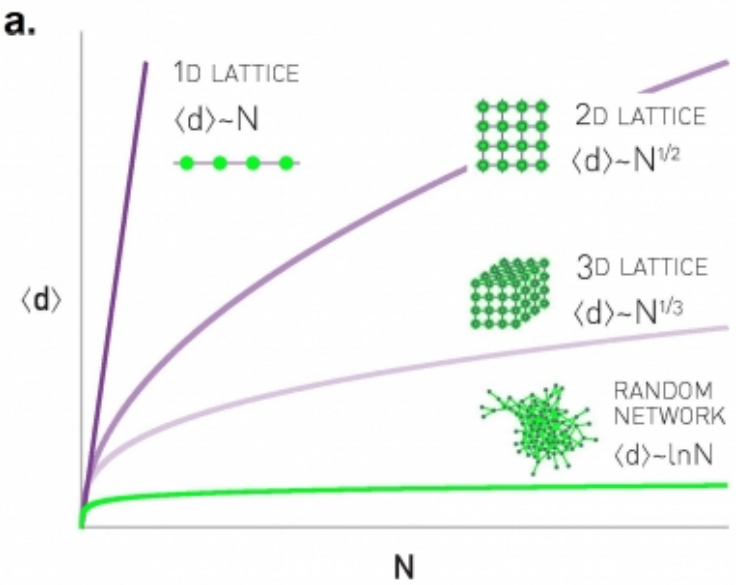
Typically, the small-world property is defined as:

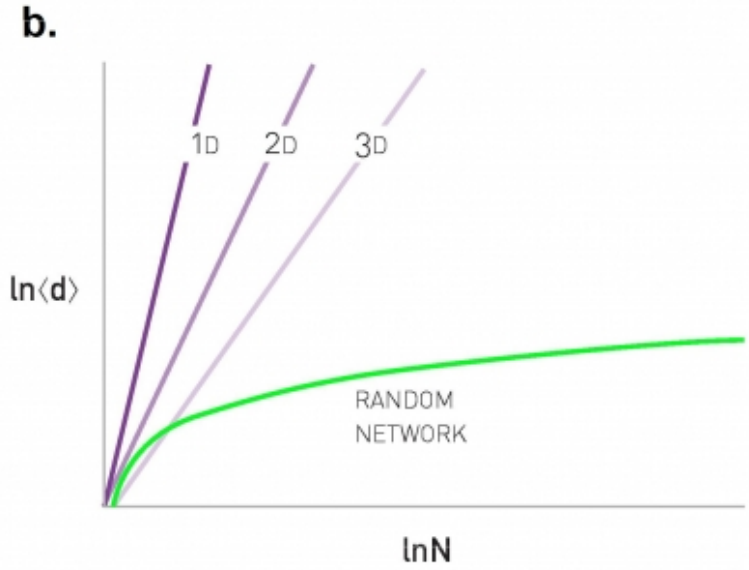
$$\langle d \rangle \approx \frac{\log N}{\log \langle k \rangle} \tag{4}$$

Short distance means that the average path length (or diameter) depends **logarithmically** on the system size.

This is a result by Manfred Kochen and Ithiel de Sola Pool (1950), which has inspired the Milgram experiment (1969).

Small-world comparing to lattice

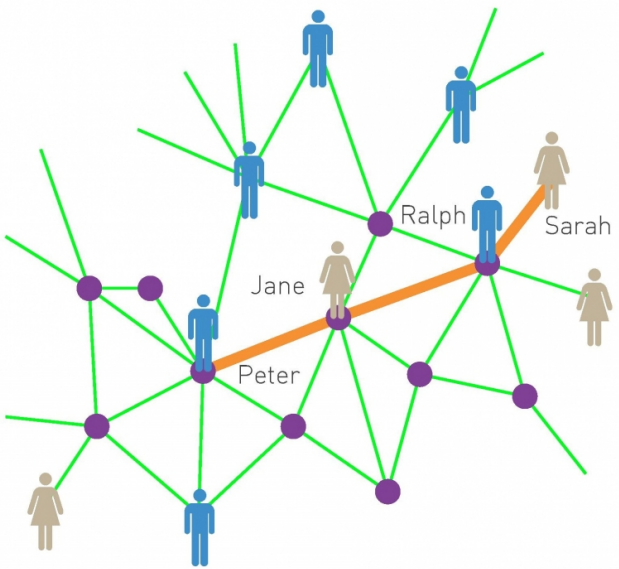




Small-world in social networks

For all individuals on Earth ($N \approx 7 \cdot 10^9$), assuming $\langle k \rangle \approx 10^3$:

$$\langle d \rangle = \frac{\log(7 \cdot 10^9)}{\log(10^3)} \approx 3.28$$



Watts and Strogatz model

[Watts and Strogatz, *Collective dynamics of 'small-world' networks* (1998)]

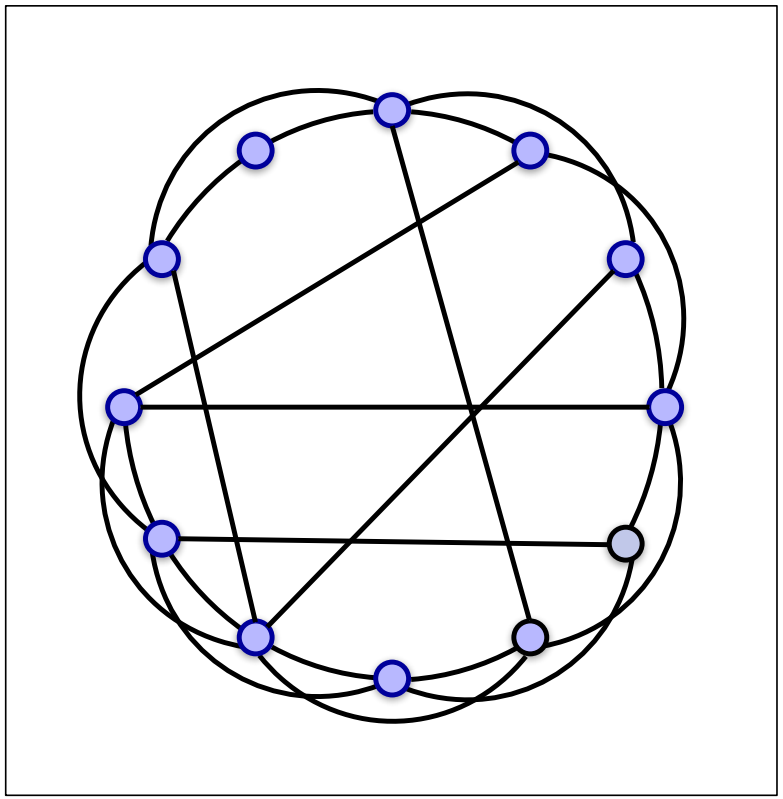
- A random graph generation model that produces graphs with small-world properties
- Starting from arbitrary topology, adding just few random links connecting distant nodes helps decrease average distance
- The resulting network features many clusters of connected people
- **This model is not intended to reproduce the growth of social networks, but shed light on the underlying structure that produce this result**

WS model: changing edge target

1. Start with regular ring lattice with N nodes and 2κ neighbours of each node
2. For each node i consider edges $(i,j), i < j$ and rewire it with certain probability p
3. *Rewiring* means replacing edge (i,j) with another edge (i,ℓ), ℓ is selected at random

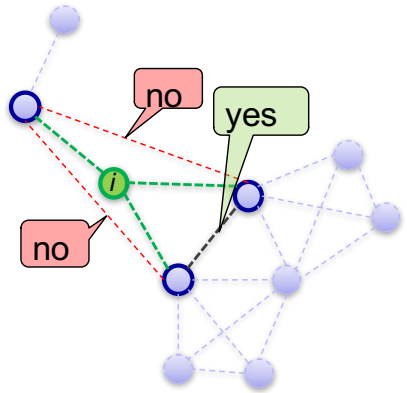
+ Degree distribution is not preserved
+ Network connectivity might break

Visual explanation



- Start from a one dimensional lattice
- Select one node at a time
- Rewire each link to a random other node with probability p
- Repeat until all it was attempted to rewire every link

Clustering coefficient

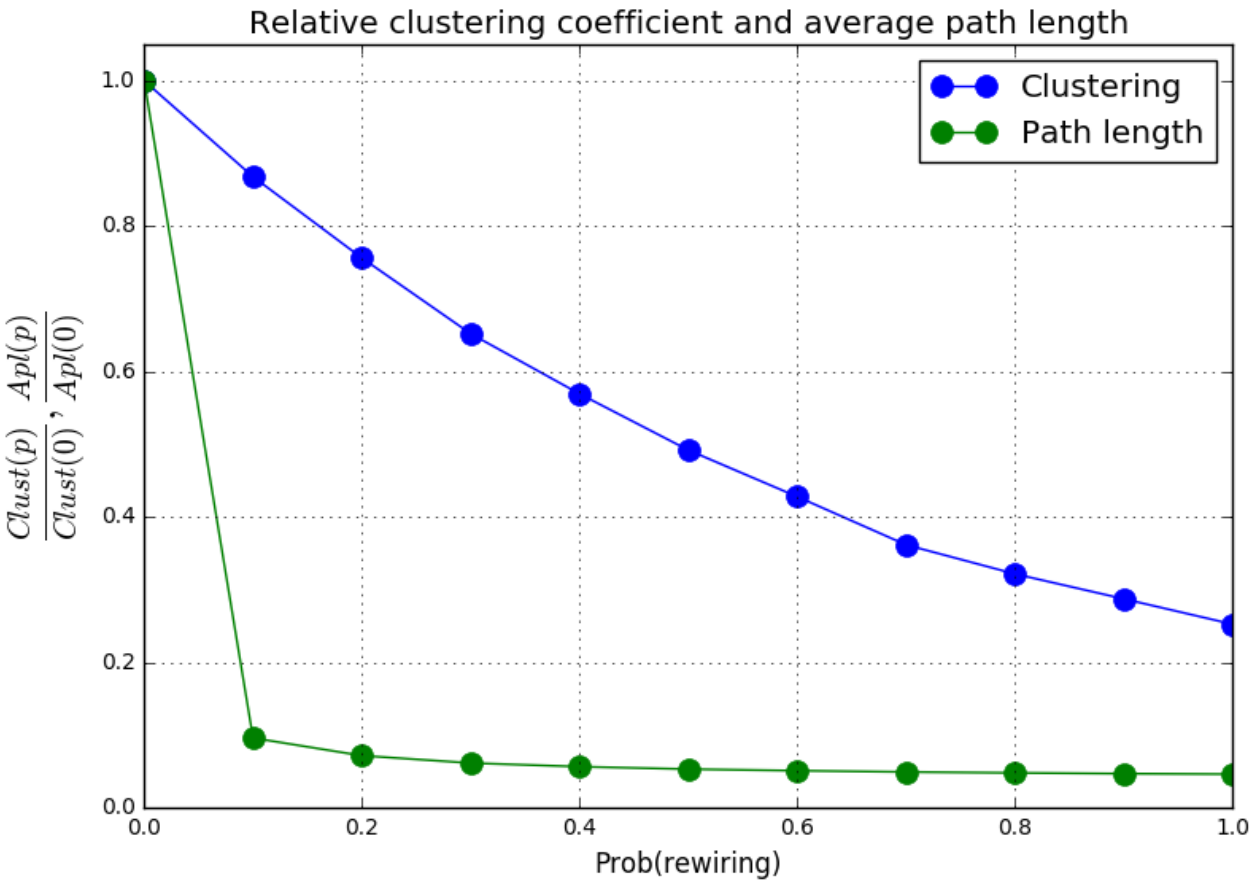


How likely is it that if two nodes are common neighbours of another, they are connected to each other?

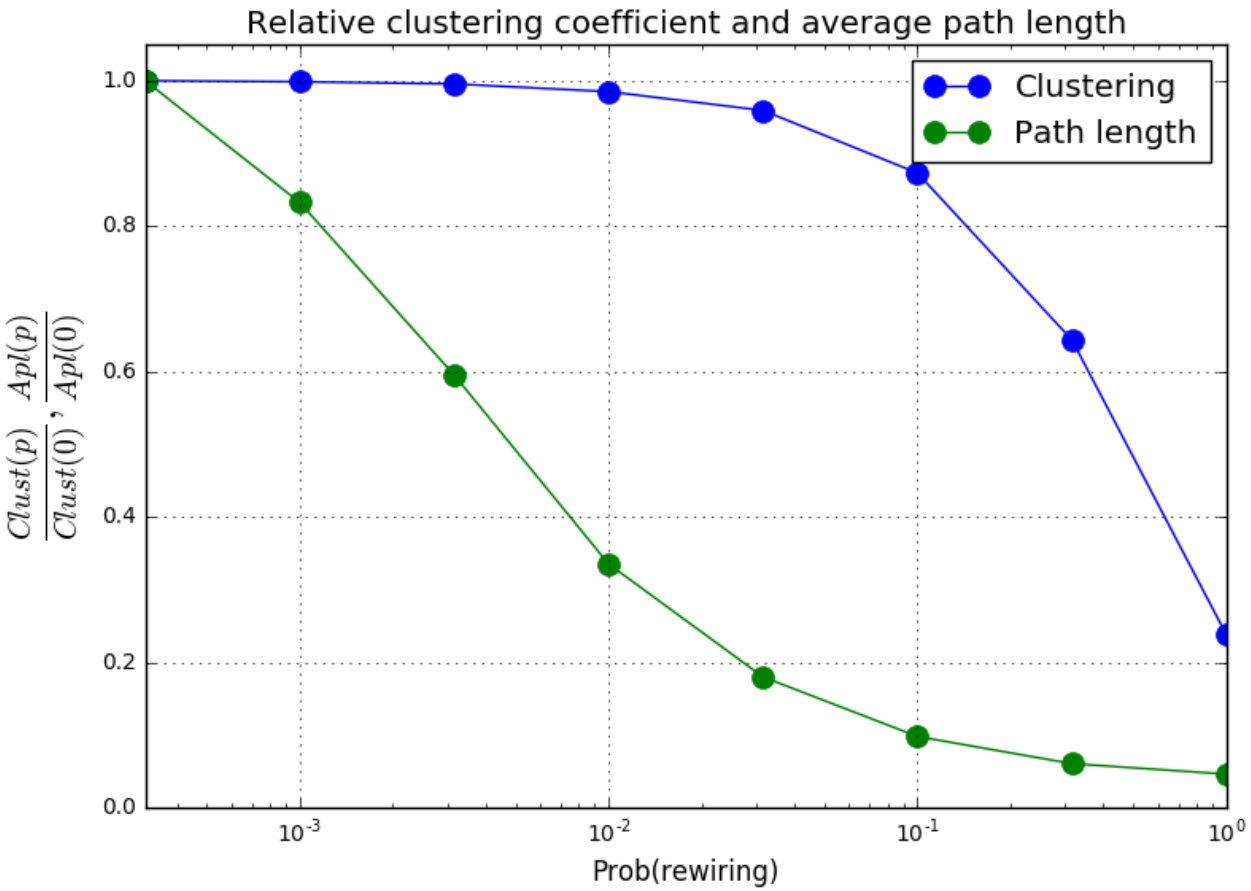
$$C(i) = \frac{\sum_{u,v \in N_i} A_{u,i} A_{v,i} A_{u,v}}{\sum_{u,v \in N_i} A_{u,i} A_{v,i}} = \frac{2 \sum_{u,v \in N_i} A_{u,v}}{k_i (k_i - 1)} \qquad (5)$$

number of closed triangles divided over number of all possible triangles

WS model: changing edge target



WS model: changing edge target



Small-world network in a nutshell

- Very few shortcuts are enough to decrease dramatically the average distance
- The relative number of necessary shortcuts vanishes for larger networks
- Average path length similar to a random network. Clustering is much larger

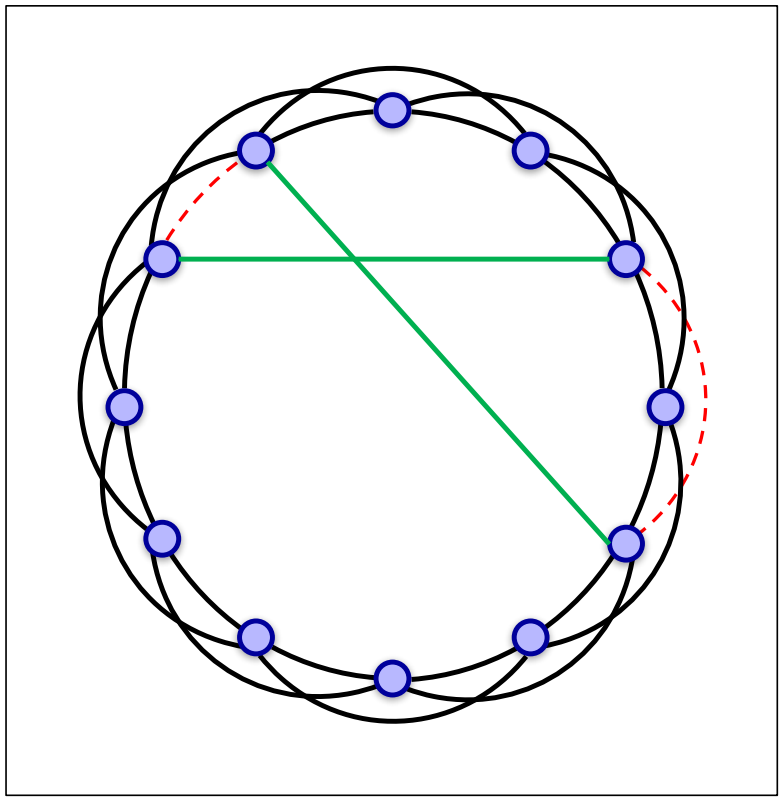
WS model: swapping edges

[Maslov, Sneppen, *Specificity and stability in topology of protein networks* (2002)]

1. Start with regular ring lattice with N nodes and 2κ neighbours of each node;
2. Select a pair of edges (i,j) and (k,l) and rewire them with certain probability p
3. *Rewiring* means replacing the edges targets, putting edges from (i,j) and (k,l) to (i,k) and (j,l).

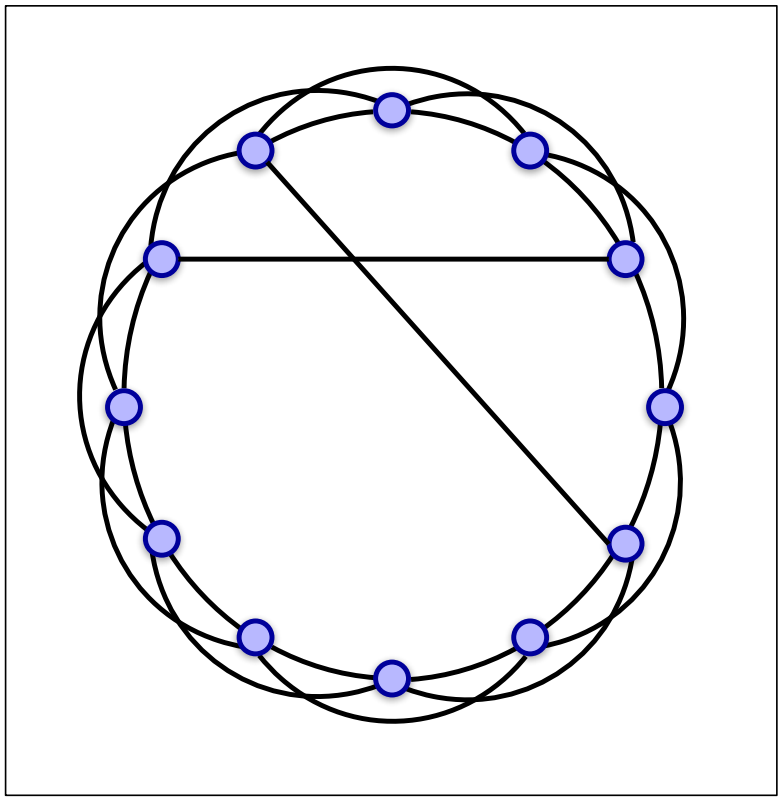
+ Degree distribution is preserved
+ Network connectivity does not break

Visual explanation



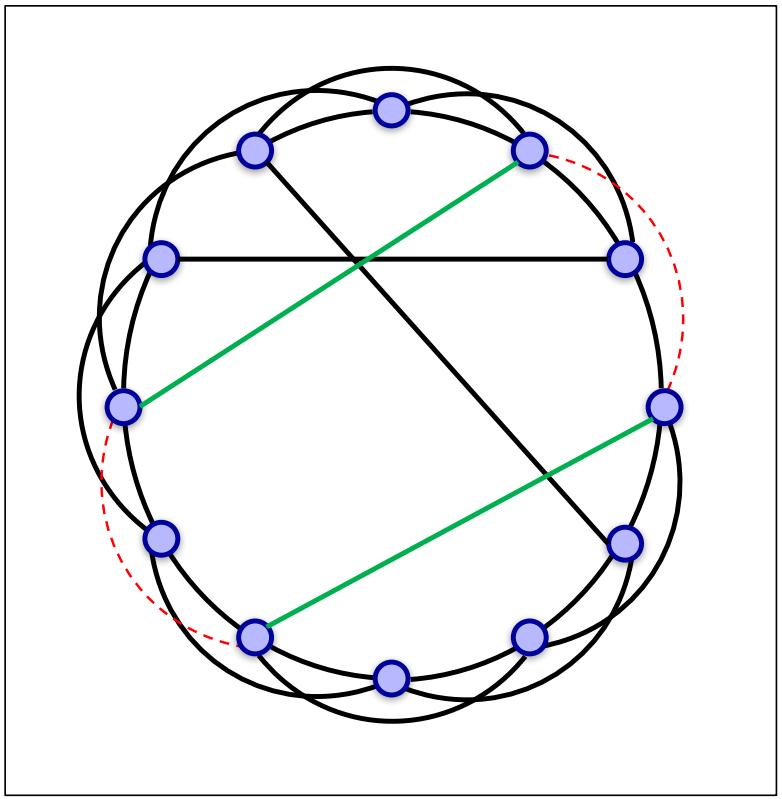
- Start from a one dimensional lattice
- Select two edges at a time
- Exchange ends with probability p
- Avoid creation of self-loops or parallel edges

Visual explanation



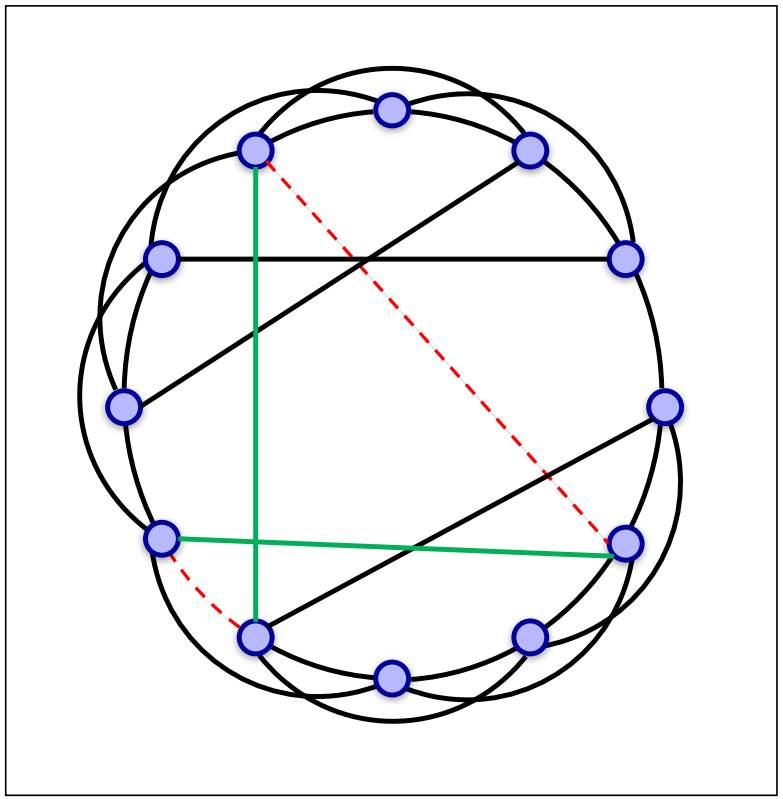
- Start from a one dimensional lattice
- Select two edges at a time
- Exchange ends with probability p
- Avoid creation of self-loops or parallel edges

Visual explanation



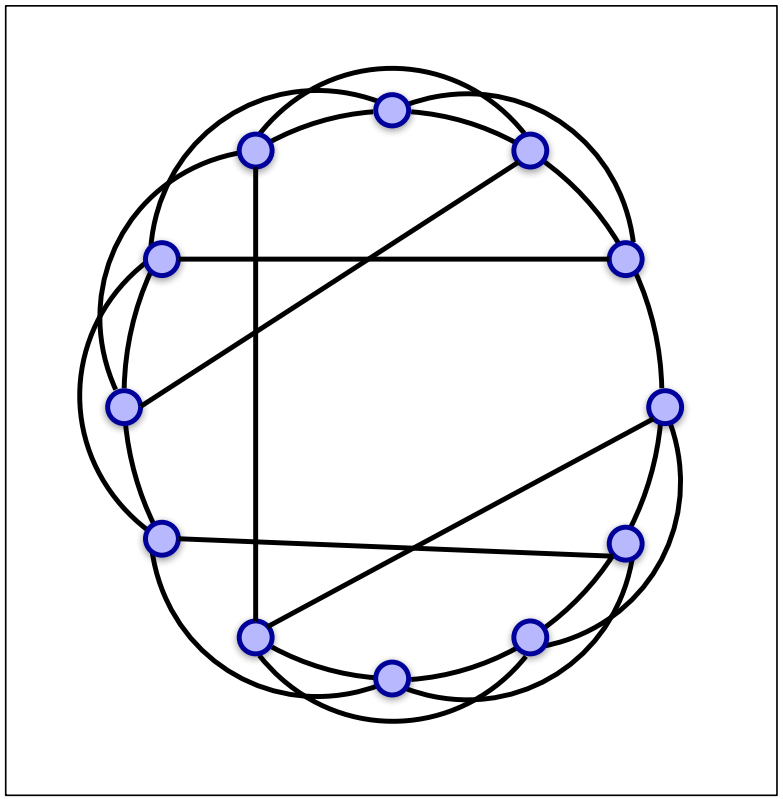
- Start from a one dimensional lattice
- Select two edges at a time
- Exchange ends with probability p
- Avoid creation of self-loops or parallel edges

Visual explanation



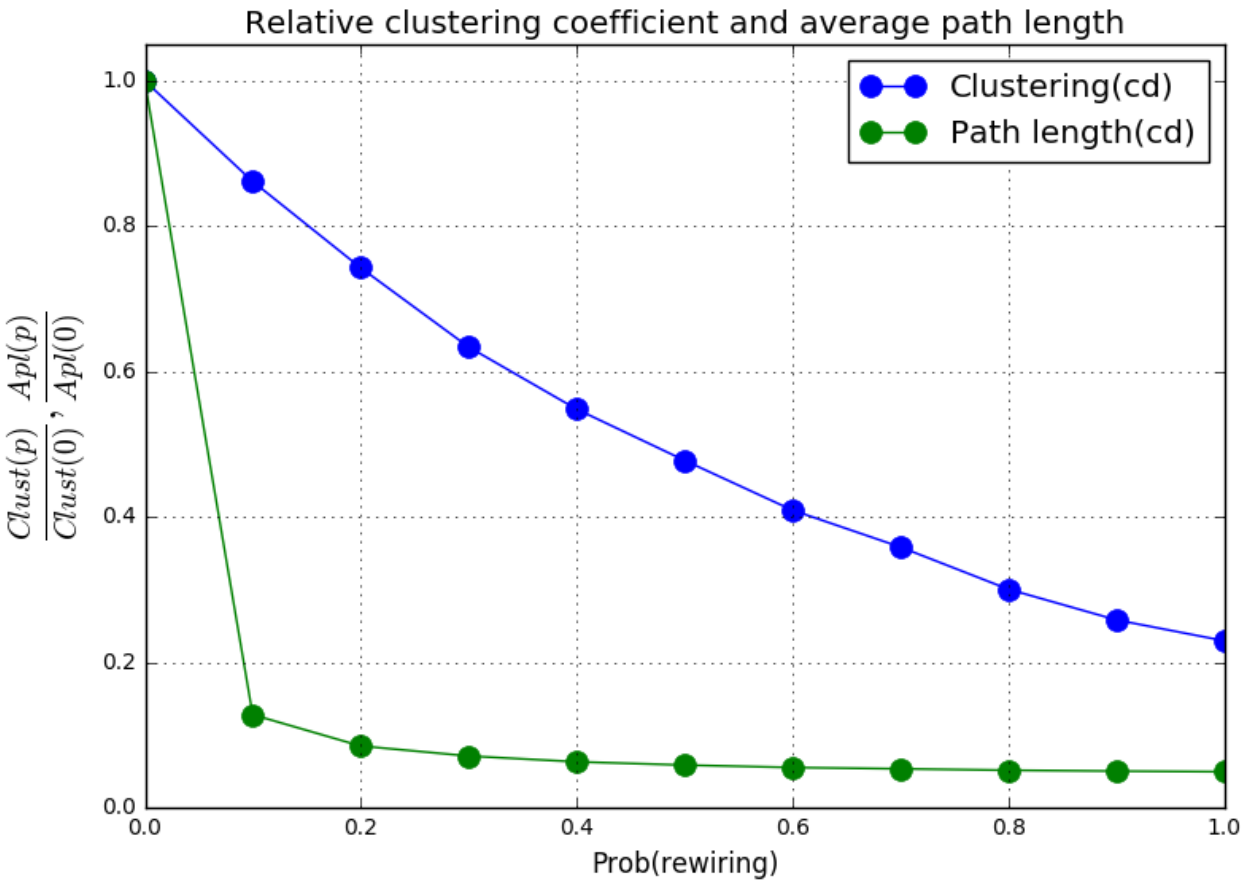
- Start from a one dimensional lattice
- Select two edges at a time
- Exchange ends with probability p
- Avoid creation of self-loops or parallel edges

Visual explanation

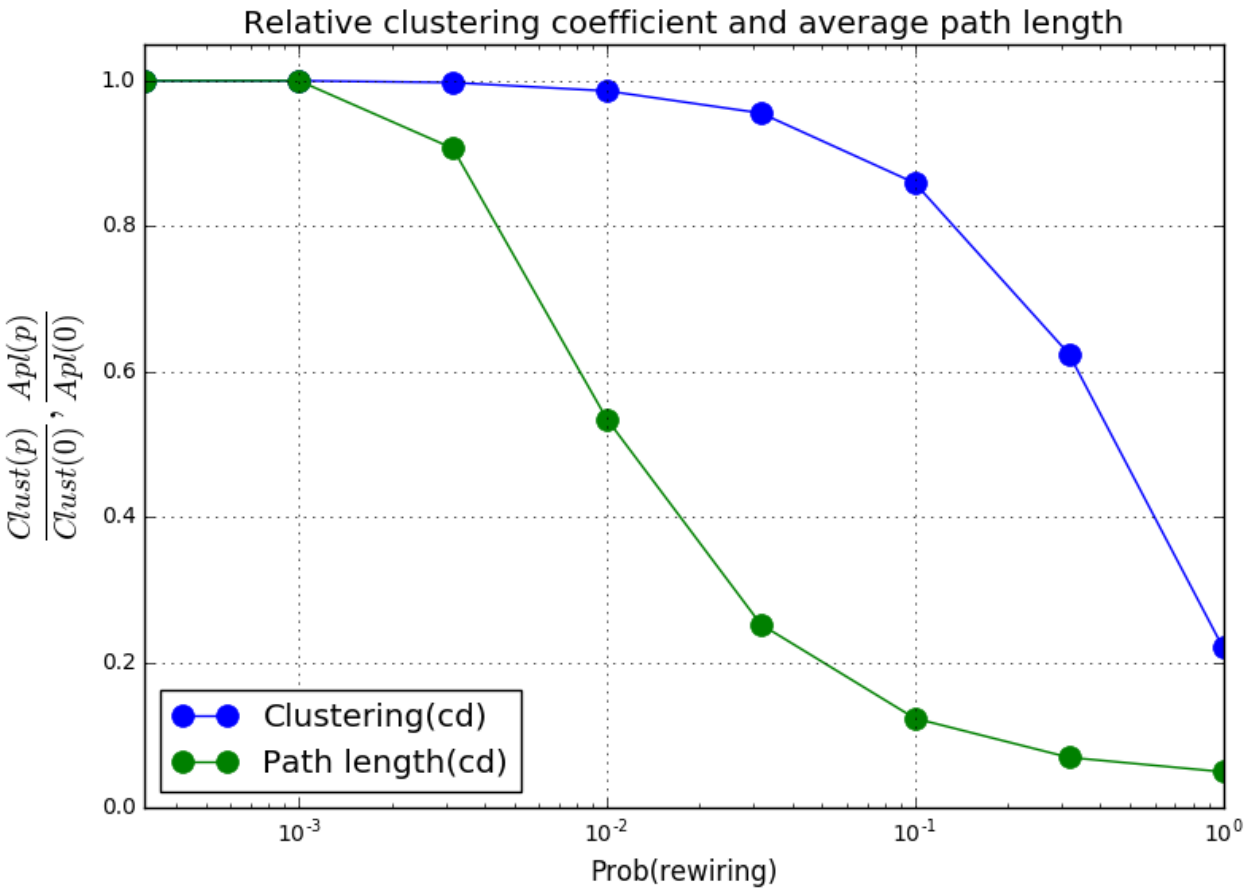


- Start from a one dimensional lattice
- Select two edges at a time
- Exchange ends with probability p
- Avoid creation of self-loops or parallel edges

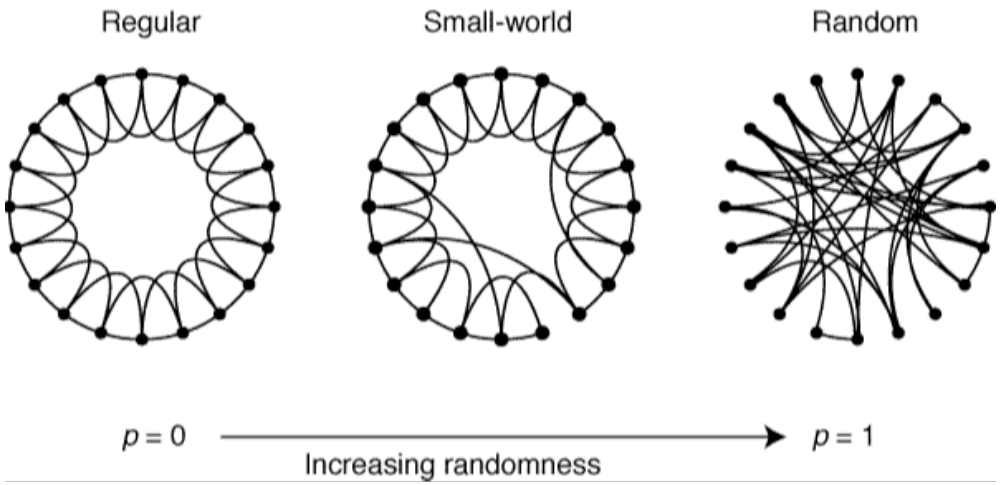
WS model: swapping edges



WS model: swapping edges



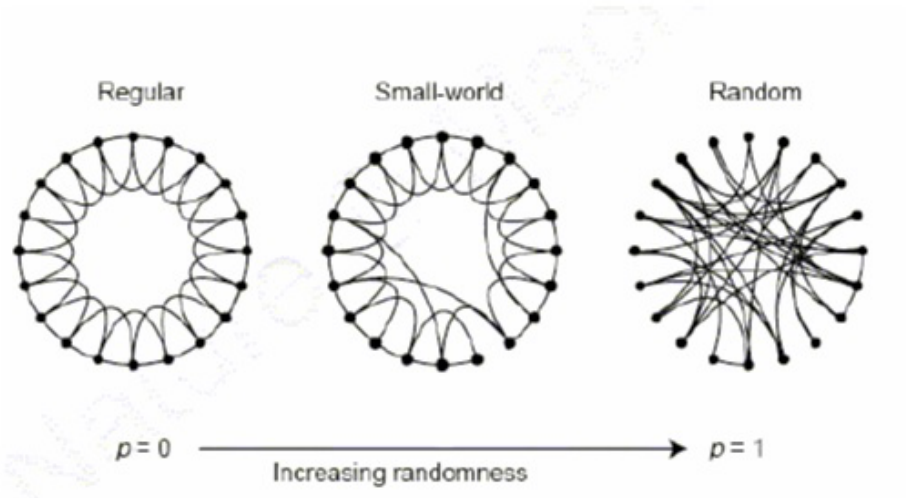
From regular to random network

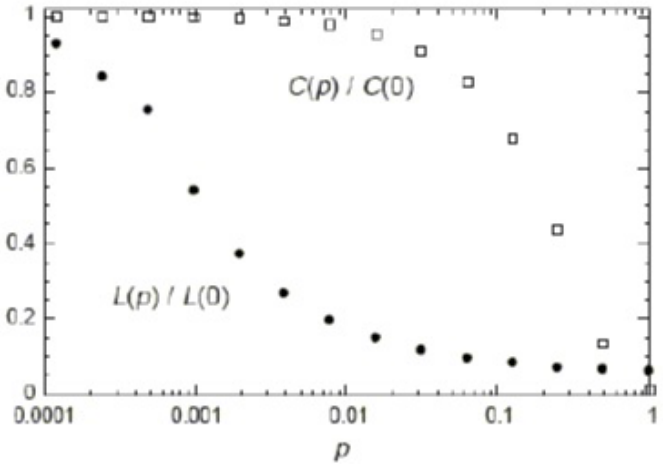


Clustering	high	high	low
Diameter	high	low	low

Small-world network summary

- Short path length (shortcuts)
- High clustering
- Unrealistic degree distribution
- Cannot be used to model network growth





References I

- Albert- Laszlo Barabasi, *Network Science*, Cambridge University Press, 2015.
- M.E.J. Newman *Networks: an Introduction*, Oxford University Press Inc., New York, 2010.
- Duncan J. Watts and Steven H. Strogatz, *Collective dynamics of 'small-world' networks*, in Nature, 1998.