7-2- Random Graphs

Complex Network Analysis Course
Reza Rezazadegan
Shiraz University, Spring 2025
https://dreamintelligent.com/complex-network-analysis-2025/

Introduction

We want network models that can accurately resemble the structure and evolution of real-world networks.
As real networks seem somewhat random, why not consider networks whose edges are spread randomly. For example in a gathering of strangers, people get to know each other nearly randomly.

In the random graph model, the probability of link formation between any two nodes is a fixed number p.
There are two models for random graphs, both introduced in 1959:

- Erdős and Rényi, $G(N, L)$: the number of nodes N and links L are fixed and the edges are spread randomly.
- Gilbert, $G(N, p)$: the number of links is not fixed and we have the probability of link formation between two nodes is p.

In the Erdős and Rényi model, the link probability (probability of link formation between two randomly chosen nodes) is $\frac{L}{\binom{N}{2}}$. For convenience we denote $\binom{N}{2}$ by $\mathcal{N}$. Note that a random graph model such as $G(N, L)$ should be thought of as an ensemble or probability distribution over the set of all graphs with N nodes and L edges.

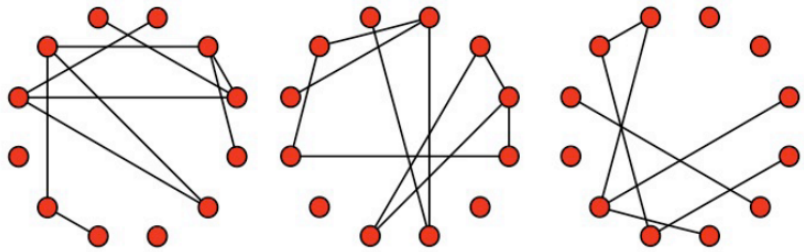
In the $G(N, p)$ model, the number of links is not fixed. Let's compute its expected value. The probability $p(L)$ of having L links is the product of three factors:

- p^L for the L links that exist,
- $(1-p)^{\mathcal{N}-L}$ for the links that do not exist,
- $\binom{\mathcal{N}}{L}$ for the number of ways to spread L links between the nodes.

Therefore $p(L)$ follows a binomial distribution and

$$E[L] = p\mathcal{N}.$$

We also have $E[k] = \frac{2E[L]}{N} = p(N-1)$ for the average degree.



Three realizations of a random graph with $N = 12$, $p = 1/6$. Image source: Barabasi, Network Science

Degree distribution

As we see in the above figure, some nodes in a random network are more connected than others, therefore we also compute the degree distribution of such a network. The probability $p(k)$ that a given node has degree k is the product of three terms:

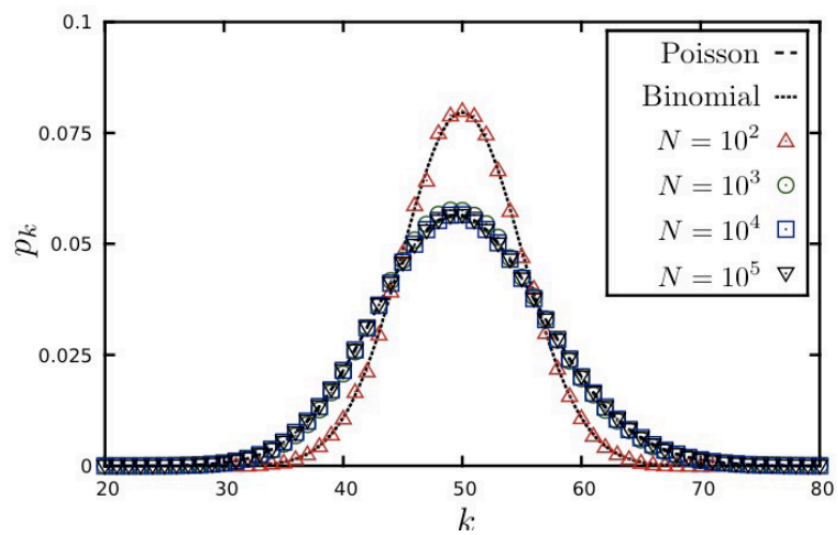
- p^k for the nodes it links to,
- $(1-p)^{N-1-k}$ for the nodes it doesn't link to,
- $\binom{N-1}{k}$ for the number of ways we can choose the linked nodes.

Thus $p(k)$ has a binomial degree as well. Most networks are sparse and therefore $E[k] = p(N-1) << N$, or in other words, p is very small. Thus, the second term above can be approximated by $e^{-E[k]}$ and the third term by $\frac{(N-1)^k}{k!}$. So, the binomial distribution can be approximated by the *Poisson distribution*:

$$p(k) = e^{-E[k]}\frac{E[k]^k}{k!}.$$

Poisson distribution: if the average number of customers visiting a shop in a day is λ, then what is the probability of k customers visiting in a given day?

Note that in this approximation, **degree distribution is independent of network size.** For example, the number of friends you have is independent of the number of people in the world!



Degree distribution for networks with average degree of 50 and network sizes 100, 1000, 10000. For $N = 100$ degree distribution varies greatly from the Poisson distribution because the condition $E[k] << N$ is not satisfied.

Clustering coefficient

CC is the probability of two neighbors of a node to be neighbors themselves. In a random graph probability of neighborhood is the same for all nodes and thus average CC equals the probability $p = E[k]/(N-1)$.

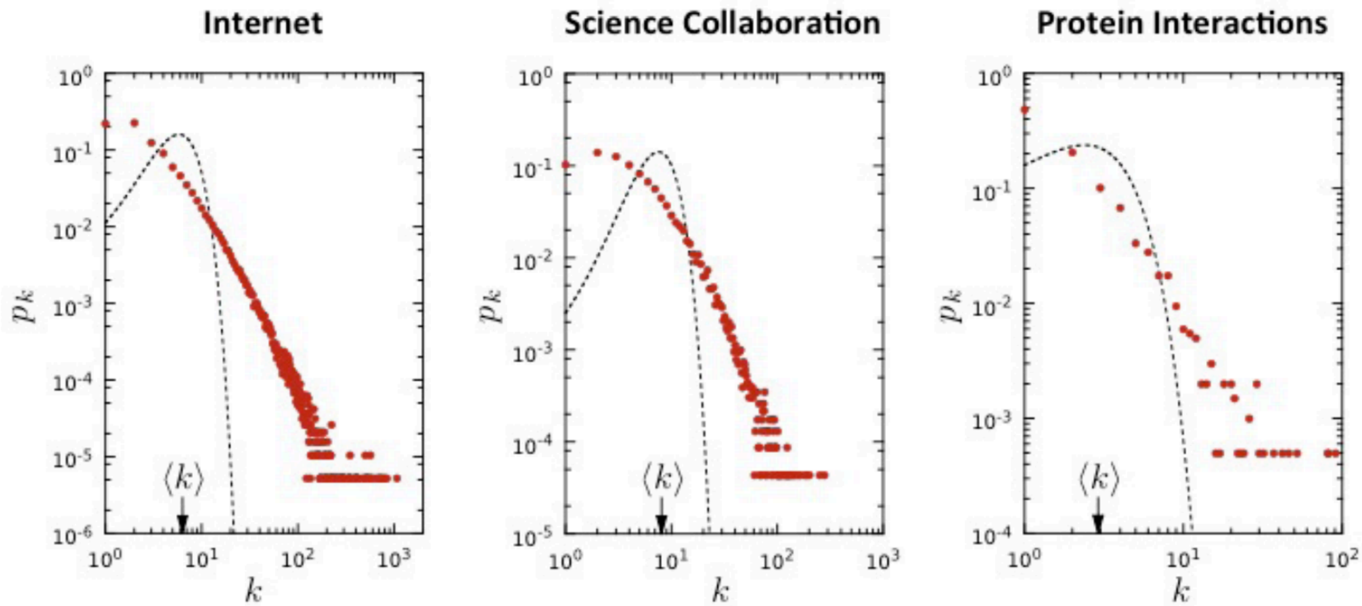
Hubs are absent in random networks

Note that in the formula for the Poisson distribution, λ^k grows slower than $k!$ and thus, the probability of having large degree decreases fast. Using the Stirling approximation for $k!$ one can show that this probability decreases super-exponentially.

Stirling approximation:

$$n! \sim \sqrt{2\pi n}(n/e)^n.$$

Poisson distribution is highly concentrated around its mean.



Comparing the degree distribution of three real networks to the Poisson distribution.

The evolution of a random graph

In this section we are concerned with the evolution of a random graph when we increase p. Extreme cases:

- $p = 0$: graph has N connected components (isolated nodes).
- $p = 1$: complete graph.

Let the *giant component* G be a component whose size is proportional to the size of the network. In other words if N_G is the size of this component then $\lim_{N\to\infty} N_G/N \neq 0$.
How does N_G/N, the relative size of the giant component evolve as we increase p? Note that we are concerned with the asymptotic behavior of the random graph when $N \to \infty$.

Note that we have two parameters in a random graph: the probability p and the number of nodes N. Increasing either one will increase the number of edges. (Note that if we add a new node to a random graph of size N then new edges are added from that node to the old nodes with the probability pN.) In the analysis of random networks, we are interested in the behavior of the network as p goes from 0 to 1 and as $N \to \infty$.

We say that the random graph $G(N, p)$ satisfies a condition (e.g. having a giant component of nonzero size for $p > 1/N$) *with high probability* if the probability of satisfying this condition (i.e. the fraction of the graphs of the form $G(N, p)$ that satisfy this property) goes to 1 as $N \to \infty$. In the following, all the properties that we consider are evaluated this way.

Erdős and Rényi showed that when we pass the threshold $E[k] = 1$ a giant component emerges. This is equivalent to $p = 1/(N-1) \simeq 1/N$.

Driving Erdős-Rényi criterion for the emergence of giant component (optional)

Let N_G be the size of the giant component. We want to compute $S = N_G/N$. Let $a = 1 - S$, be the fraction of the nodes not in GC. Every node in GC is connected to another node in CG.
To compute the probability a, note that if u is a node not in GC, and v is another node, then either:

- v is in GC and u is not connected to it, with probability: $(1-a)(1-p)$
- v is not in GC, with probability: a.

Thus, the probability that u is not connected to GC via v is $(1-a)(1-p) + a = 1 - p + ap$.
The probability that u is not connected to GC via any node: $(1-p+ap)^{N-1}$. Thus we have:

$$a = (1 - p + ap)^{N-1}.$$

Since $E[k] = p(N-1)$,

$$a = \left(1 + E[k]\frac{a-1}{N-1}\right)^{N-1} \to e^{E[k](a-1)}.$$

This is because $E[k] << N$ and we are interested in the asymptotic behavior when $N \to \infty$.
Thus

$$a \simeq e^{-E[k](1-a)}.$$

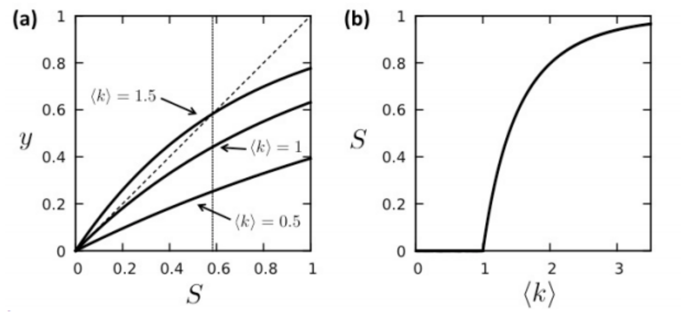
If $S = N_G/N = 1 - a$ then

$$S = 1 - e^{-SE[k]}.$$

Exercise: show that if $E[k] < 1$ and $S > 0$ then $1 - e^{-SE[k]} < S$. Thus S has to be zero.

It follows from the exercise that when $E[k] < 1$, $S = 0$ and thus we don't have a giant component (i.e. the size of the giant component is zero).

Exercise: show that if $E[k] > 1$ then the equation for S has a nonzero solution and thus we have a giant component. Hint: show that for different values of S, $1 - e^{-SE[k]} - S$ can be both positive and negative.



Left: plot of $y = 1 - e^{SE[k]}$ and its intersection with $y = S$. Right the plot of S as a function of $E[k]$. Source: Barabasi, Network Science

Note that at the turning point, where the plot hits the $y = S$ line, the derivative of $1 - e^{-SE[k]}$ w.r.t. S equals 1, which gives $E[k]e^{-SE[k]} = 1$ or $\log(E[k]) + (-SE[k]) = 0$ or $S = \log(E[k])/E[k]$. Since S is nonnegative, this means that the transition from $S = 0$ to $S > 0$ happens at $E[k] = 1$.

What this means is that if $E[k] < 1$, the fraction of the elements of the random graph which belong to any given component goes to zero as $N \to \infty$.

Exercise: show that the probability of the existence of two giant components decreases exponentially as $N \to \infty$. Hint: assume we have two giant components and compute the probability that no edges exist between the two.

We call the components other than giant component as *small components*. If follows from the above discussion that if s is the size of a small component then $s/N \to 0$ as $N \to \infty$.

Computing the sizes of small components (optional)

Regardless of the value of $E[k]$, we show that small components are trees and compute the distribution of their sizes. More precisely we show that "with high probability" (as defined above), small components are trees.

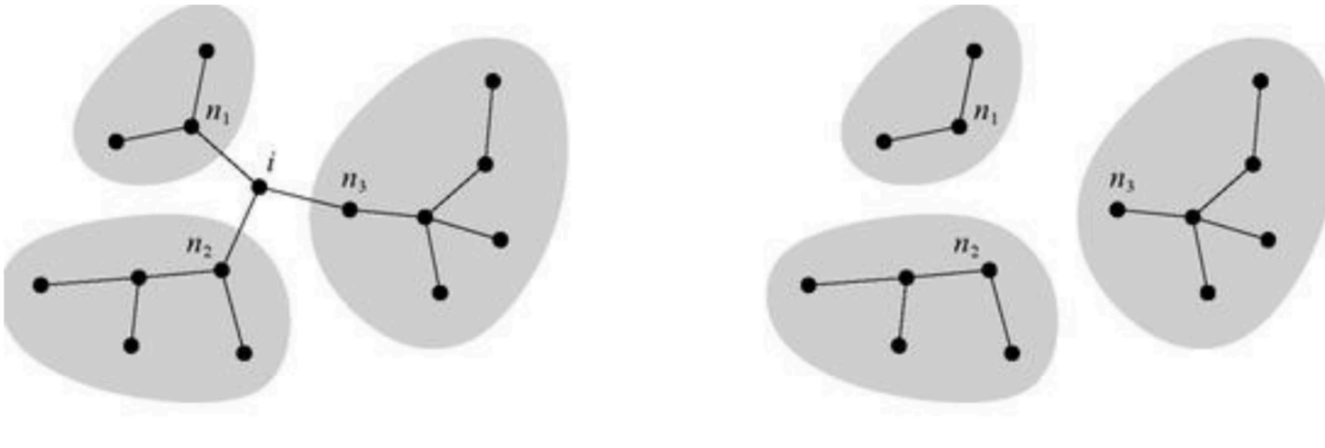
If a component of size s is not a tree then it contains an edge between two nodes such that if we remove this edge, the component stays connected. The probability of having such an edge is

$$[\binom{s}{2} - (s-1)]p.$$

The second term in the brackets is to exclude vertices that are already connected. The above probability equals $(s-1)(s-2)\frac{E[k]}{N-1}$. Thus, if we show that $\lim_{N\to\infty} s^2/N \to 0$ then this probability goes to zero.

Let π_s be the probability that a randomly chosen vertex belongs to a small component of size s. Note that $\sum_{s=1}^{\infty} \pi_s = 1 - S$ where S is the size of the giant component.

The probability $p(s|k)$, that such a vertex has degree k too, can be computed by realizing that the removal of this vertex, decomposes the tree into a set of k trees whose sizes add up to $s-1$.



(a) (b)

We have:

$$p(s|k) = \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \prod_{j=1}^{k} \pi_{s_j} \cdot \delta(s-1, \sum s_j)$$

where δ is the Dirac delta function. For example $p(s|2) = \sum_{s_1=1}^{\infty} \sum_{s_2=1}^{\infty} \pi_{s_1}\pi_{s_2}\delta(s-1, s_1+s_2)$.
Of course: $\pi_s = \sum_k p(s|k)p(k)$ and $p(k)$ is the degree distribution given by Poisson distribution.
We consider the *generating function* $h(z) = \sum_s \pi_s z^s$.

Exercise: use the recursion formula above to show that $h(z) = z\exp(E[k](h(z)-1))$.

Note that $h(1) = 1 - N_G/N = 1 - S$ and $E[s] = \frac{\sum_s s\pi_s}{\sum_s \pi_s} = \frac{h'(1)}{1-S}$.

Exercise: using the last exercise show that $h'(1) = \frac{1-S}{1-E[k]+SE[k]}$.

Thus the average size of a small component to which a randomly chosen vertex belongs is $E[s] = \frac{1}{1-E[k]+SE[k]}$. Note that when $E[k] < 1$ then $S = 0$. As $E[k]$ approaches 1, the average size of components goes to infinity. It becomes finite again for $E[k] > 1$.

Note that if n_s is the number of components of size s then $\pi_s = sn_s/N$.
Exercise: show that $E[n] = 2/(2 - E[k] - SE[k])$.

Using the theory of complex variables, it can be shown that

$$\pi_s = \frac{(sE[k])^{s-1}}{s!} e^{-sE[k]}.$$

Using Stirling formula we get

$$\pi_s \sim s^{-3/2}\, e^{s(1-E[k])}\, E[k]^{s-1}.$$

Subcritical level $E[k] < 1$ or $p < 1/N$

The number of edges is nearly $E[L] = pN(N-1)/2 < N/2$.
As we saw above, in this case we don't have a giant component and thus, the fraction of nodes belonging to any component goes to 0, in the $N \to \infty$ limit.

The connected components are trees. The size of the largest connected component is $\log N$ and thus the ratio N_G/N goes to zero for large N. (The size of the largest connected component can be defined by $\sum_{s=s_{max}}^{\infty} \pi_s = s_{max}/N$.)

When average degree is less than 1, the last terms in π_s dominate, and thus, the probability of having a large component of size s decreases exponentially with s.

Critical level $E[k] = 1$

When $E[k] = 1$, in equation for π_s we get $p(s) \sim s^{-3/2}$.
The size of the largest component is $N^{2/3}$. Most nodes are in various small components.

Supercritical level $E[k] > 1$

$$S = \frac{N_G}{N} \sim E[k] - 1.$$

Connected level $E[k] \geq \log N$ or $p \geq \log N/N$

In this regime, as $N \to \infty$, the size of the giant component goes to N. From the equations above this means that $pN \to \infty$.

Exercise: show that in the formula for $a = 1 - S$, in terms of N, if $E[k] = \log N$ or is larger, then if the right hand side is convergent then it goes to zero, which means $a = 0$ or $S = 1$.

Network is still relatively sparse.

Real networks are supercritical

Network	N	L	<k>	ln N
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	186,936	8.08	10.04
Actor Network	212,250	3,054,278	28.78	12.27
Yeast Protein Interactions	2,018	2,930	2.90	7.61

The size and average degree of a few real networks. Source: Barabasi, Network Science.

Even though these networks are supercritical (thus they have a giant component according the random graph model), for most of them $E[k] < \log N$ and thus according to random graph theory they must not be connected, which is a contradiction.

Exercise: a) Using the Networkx library, simulate the behavior of a random network with 100 nodes as p goes from 0 to 1. Consider the values 1/1000, 1/500, 1/300, 1/200, 1/100, 1/90, 1/80, 1/50, 1/30, 1/20, 1/10 for p. For each value plot the network and plot the distribution of the sizes of the components.
b) For random networks of sizes 100, 1000, 10000, for critical, a subcritical and a supercritical value of p compute the fraction of the nodes belonging to the largest component.

Small world property

Let d denote shortest path distance in the network. Assuming d to be small, the number of nodes at distance d of a node is proportional to $E[k]^d$.
Thus the number $N(d)$ of nodes within distance d of a given node can be approximated by

$$1 + E[k] + E[k]^2 + \cdots + E[k]^d = \frac{E[k]^{d+1} - 1}{E[k] - 1}.$$

If $E[k] >> 1$, this number can be approximated by $E[k]^d$.

If d_{max} is the diameter of the network, from $N(d_{max}) = N$ we get a rough approximation:

$$d_{max} \propto \frac{\log N}{\log E[k]}.$$

For most networks this equation provides a better approximation for average distance than d_{max}, because the latter is dominated by some outlying nodes. Thus small world property can be defined by

$$E[d] \propto \frac{\log N}{\log E[k]}$$

Note that $\log N$ is orders of magnitude smaller than N, thus average distence will be surprisingly small.

For the global social network with $N = 7 \times 10^9$ and $E[k]$ being equal to the Dumbar number 150, we get the approximation 4.5 for the average distance.

Network Name	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\log N}{\log \langle k \rangle}$
Internet	192,244	609,066	6.34	6.98	26	6.59
WWW	325,729	1,497,134	4.60	11.27	93	8.32
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	186,936	8.08	5.35	15	4.81
Actor Network	212,250	3,054,278	28.78	-	-	-
Citation Network	449,673	4,707,958	10.47	11.21	42	5.55
E Coli Metabolism	1,039	5,802	5.84	2.98	8	4.04
Yeast Protein Interactions	2,018	2,930	2.90	5.61	14	7.14

Average distance and diameter in some real networks and their approximation based on random network theory. Source: Barabasi, Network Science.

Generating networks with a prescribed degree distribution

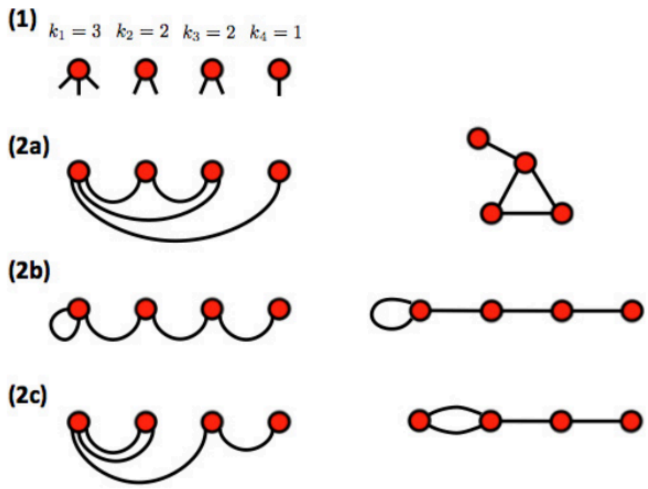
Given a degree distribution $p(k)$ we want to generate a network whose degree distribution equals $p(k)$ but is otherwise random.

Configuration model

In this model, the number of links L, as well as the number of nodes N is fixed. We start with a degree *sequence* i.e. numbers $k_1, k_2, \ldots, k_N$ which give the degrees of the nodes. We regard a node of degree k_i as having k_i "stubs" attached to it. We then randomly attach the stubs to each other.
Note that $\sum_i k_i$ has to be even, and it equals $2L$. Note also that this method can produce loops and multiple edges.

The probability of having an edge between nodes of degrees k_i, k_j equals

$$\frac{k_i k_j}{2L-1}.$$

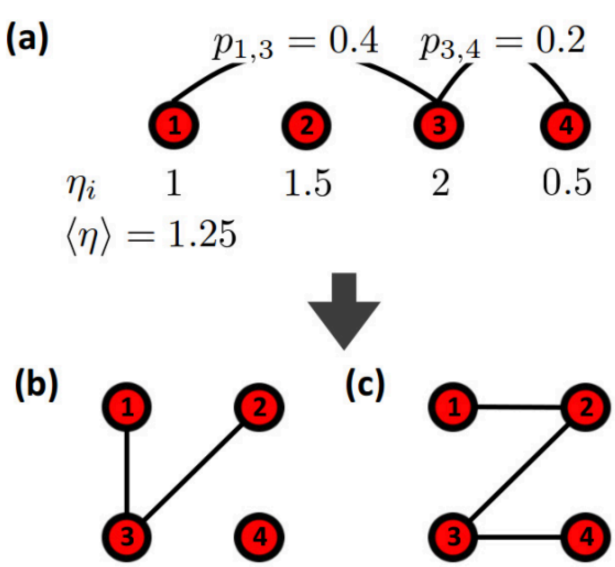


Exercise: show that as $N \to \infty$, the expected number of self-edges and multi-edges goes to zero.

Hidden parameter model

This is another model for generating networks with a given degree distribution.
Given a distribution η, and the number N of nodes, we draw N random numbers $\eta_1, \eta_2, \ldots, \eta_N$ from the distribution. These numbers represent the *fitness* of the nodes. Starting from N isolated nodes, we connect edges i, j with the probability

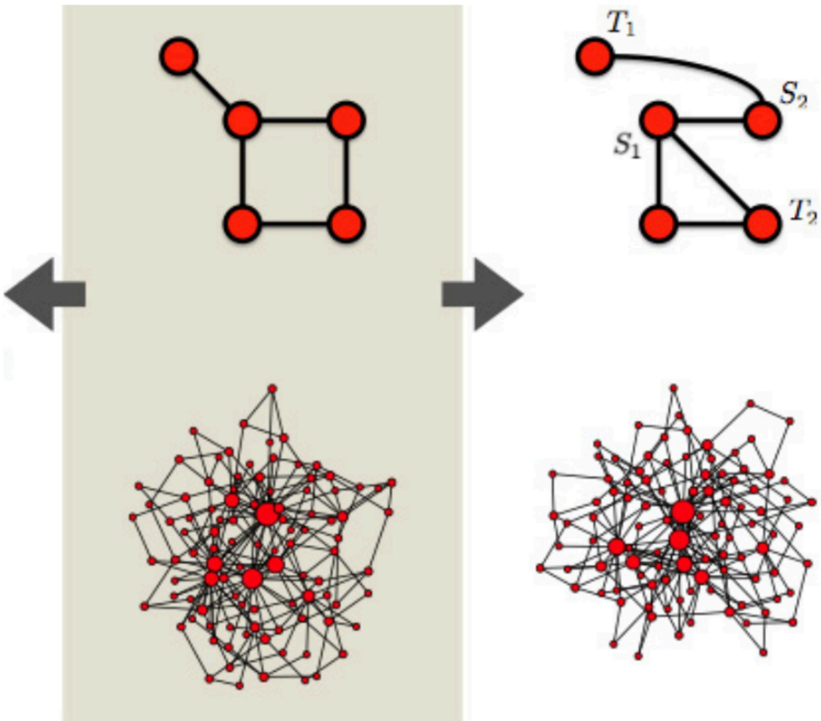
$$p_{i,j} = \frac{\eta_i \eta_j}{E[\eta] N}.$$



Exercise: a) What is $p(k_i = k)$?
b) What is the degree distribution of the resulting network?

Degree preserving randomization

In this method we "rewire" network's links in such a way that the degree of each node is preserved.



Question: is a certain network property predicted by its degree distribution alone, or does it represent some additional property not contained in $p(k)$? If the property still holds after degree preserving randomization, then it is determined by the degree distribution.

Algorithm: randomly choose two links $(u, v), (w, x)$ in the network and swap them to $(u, x), (w, v)$. The degrees of the two nodes are preserved. Do this until all links in the network are permuted.