Analytical results for the Distribution of Shortest Path Lengths in random networks

Eytan Katzav

The Racah Institute of Physics

The Hebrew University of Jerusalem

In collaboration with

Mor Nitzan, Daniel ben-Avraham, P.L. Krapivsky, Reimer Kühn, Nathan Ross and Ofer Biham – arXiv:1504.00754 – accepted to EPL

Oldenburg, August 5 2015

Outline

- Introduction – Random Graphs, Erdos-Renyi etc..

  Shortest Paths – definition, known results
- Mathematical preliminaries
- Analytical approach:

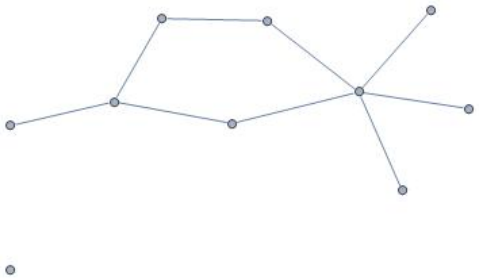
  The Recursive Paths Approach (RPA)

  Comparison with Simulations
- Dense Limit
- Local-Global correlations – conditioning on the degree of a node
- Summary and Outlook

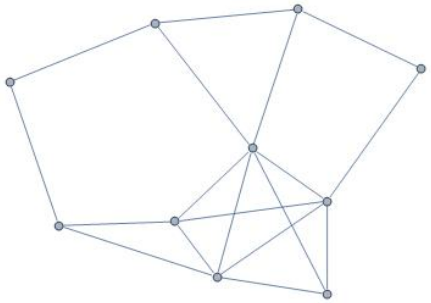
Random Graphs/Networks

Graphs are essentially a collection of vertices (nodes) and edges (links). One could think of weighted graphs, where the edges are weighted according to some rule.

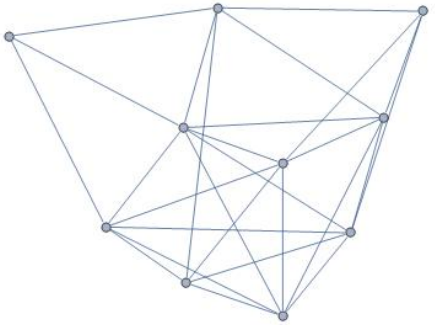
There are many interesting families of random graphs, examples being – Erdos-Renyi (ER) graphs (N nodes, and probability p for the existence of every possible edge independently), e.g. for $N = 10$



p = 0.2



p = 0.4

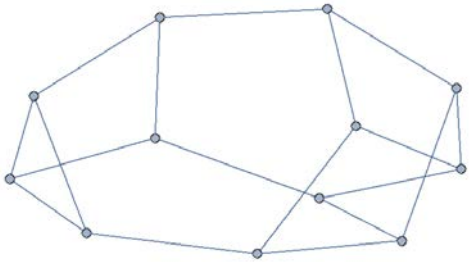


p = 0.6

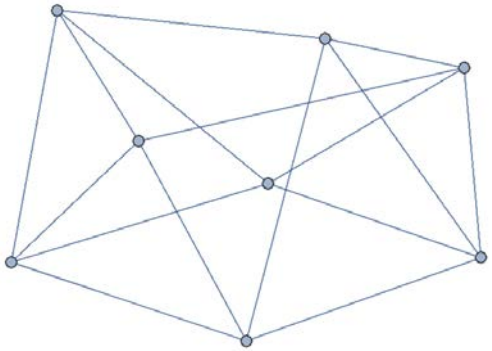
Binomial (Poisson) degree distribution $P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$

Random Graphs/Networks

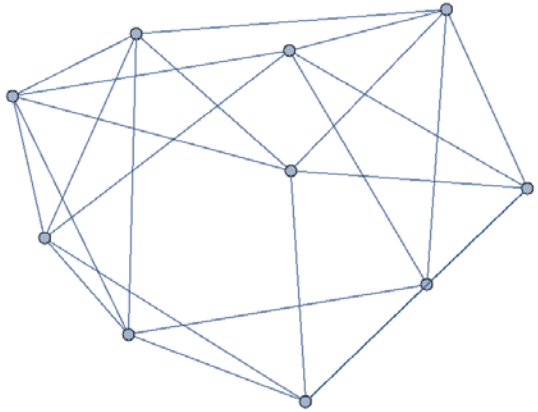
Random regular graphs (N nodes, all vertices of degree k_0, and a uniform sampling out this ensemble), e.g. $\mathrm{N} = 8$, $P(k) = \delta_{k,k_0}$



$k_0 = \mathbf{3}$
(cubic)



$k_0 = 4$

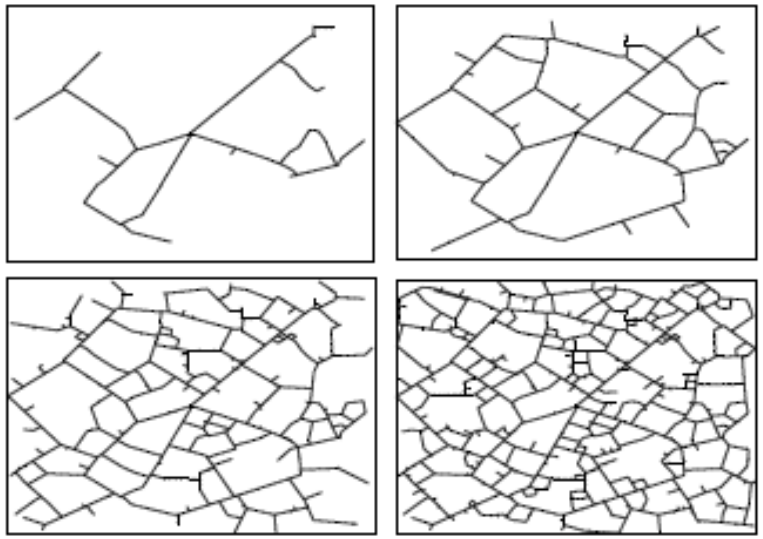


$k_0 = \mathbf{5}$

Such graphs often enjoy many analytical results

Random Graphs/Networks

Random planar graphs (random cities)



Barthelemy & Flammini 2008 - Modeling urban street patterns

Snapshots of the network at different times of its evolution. At short times, we have almost a tree structure and loops appear for larger density values obtained at larger times.

Such graphs are not always easy to generate nor to sample uniformly from, but they model nicely spatial networks such as transportation, road networks in cities, vein networks in leaves, spread of diseases.

Why are they interesting?

As **null models for real networks**:
Annibale, Coolen et al. 2010, Tailored graph ensembles as proxies or null models for real networks – suppose you have a real network with a given degree distribution, and you observe a certain pattern. A meaningful way to test for the robustness of this observation would be to compare it to a random null-model or a benchmark.

Insight into generic properties of certain networks, such as the small-world property of random scale-free networks:
In models of scale-free random networks it is possible to demonstrate that Dunbar's number (cognitive limit to the number of people with whom one can maintain stable social relationships, i.e. 100-200) is the cause of the phenomenon known as the **'six degrees of separation'** (everyone and everything is six or fewer steps away, by way of introduction, from any other person in the world).

Preliminaries

This presentation focuses on Erdos-Renyi graphs on N nodes and with probability p for every possible edge to exists independently.

We denote this ensemble by $ER(N,p)$.

A short review of some properties of ER graphs as a function of p, or the mean connectivity now follows:

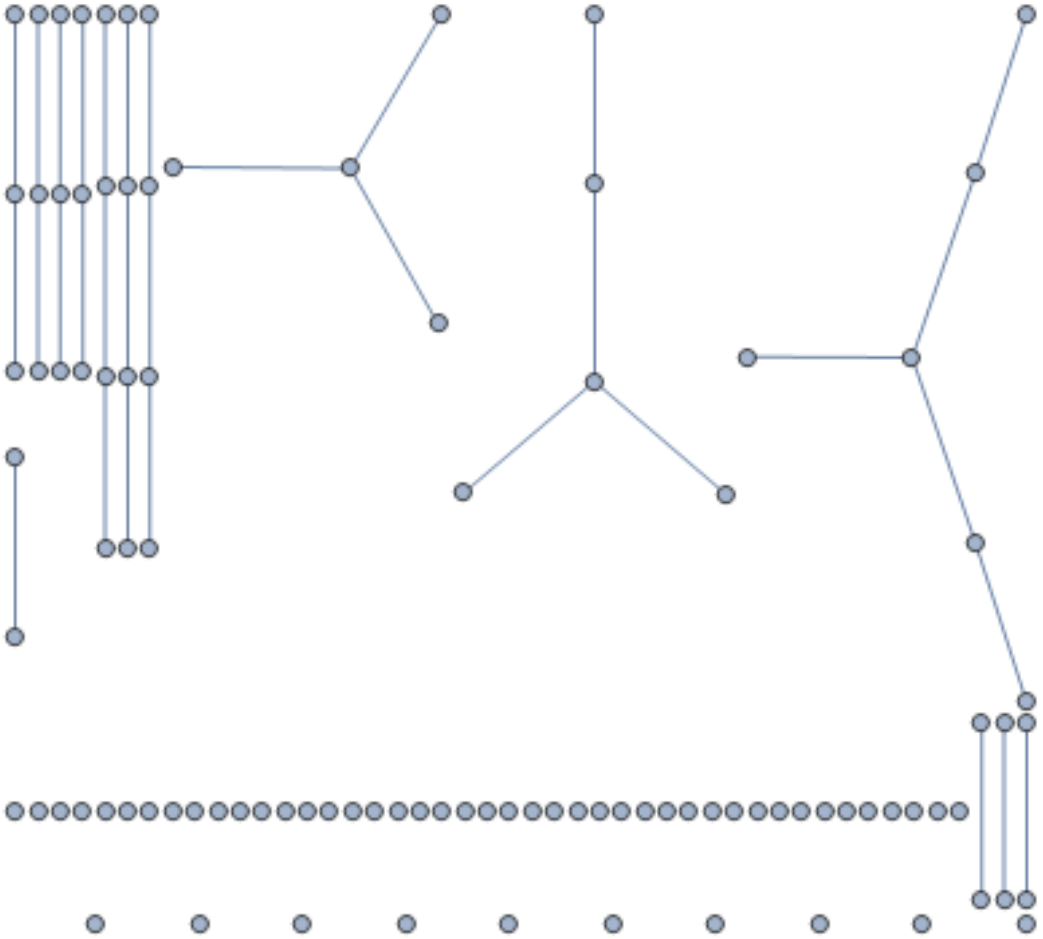
1. Sparse case – $\mathrm{p} = c/N$ where c is the mean connectivity which is of order unity in this case. There is a percolation transition at c=1.
2. Dense case – when $\langle c \rangle \geq \log \mathrm{N}$ and so there are no isolated components in the network. Important examples are

   $\langle c \rangle \sim \log \mathrm{N}$, $\langle c \rangle \sim \mathrm{N}^{\alpha}$ $0 < \alpha < 1$ (both with vanishing p)

   and $\langle c \rangle \sim \mathrm{N}$ (with p of order unity)

Examples – $\mathrm{ER}(\mathrm{N}, \mathrm{p}),\ \langle c \rangle = 1/2$

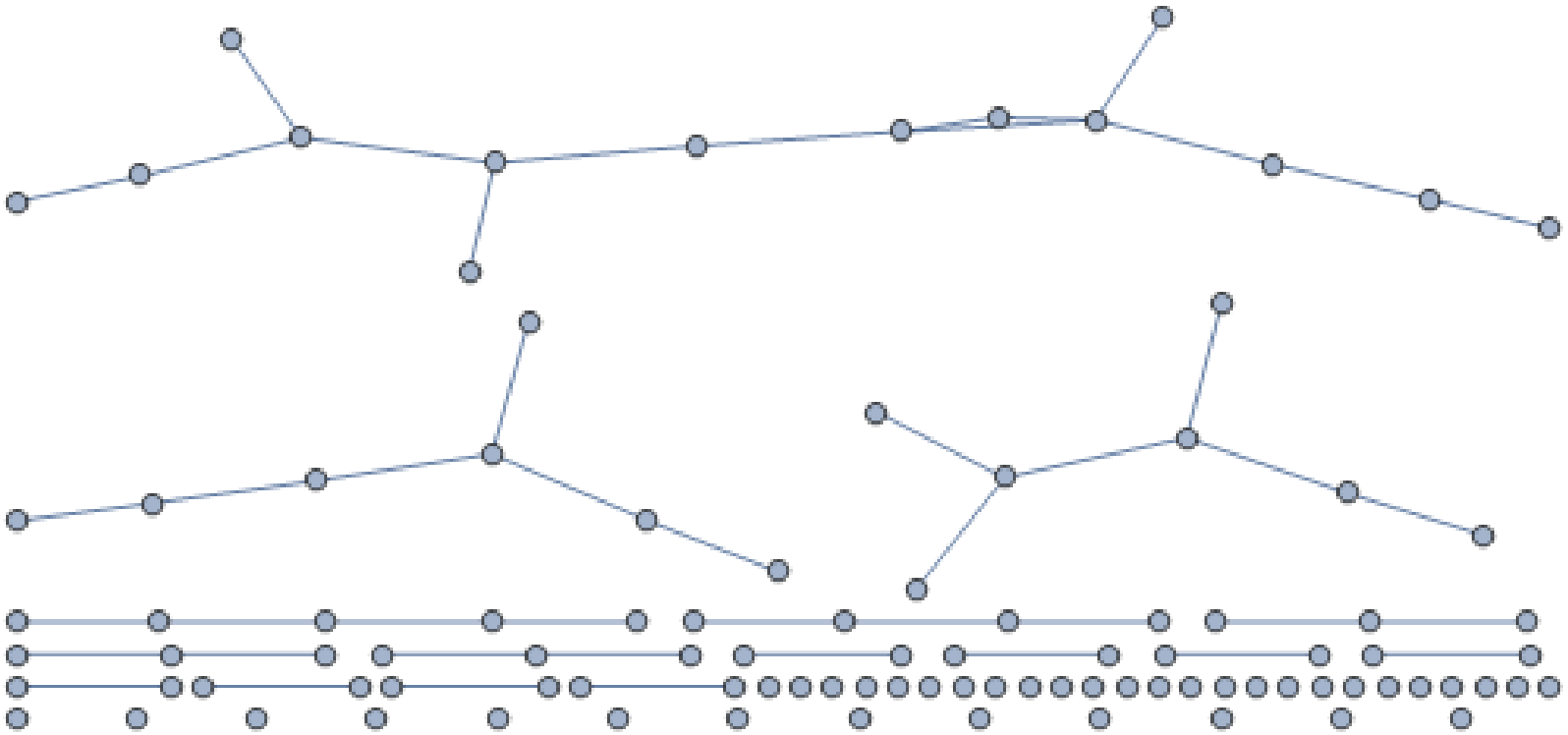
$\mathrm{N} = 100, \langle c \rangle = 1/2$ – mostly isolated components



Examples – $\mathrm{ER}(\mathrm{N}, \mathrm{p}),\ \langle c \rangle = 1$

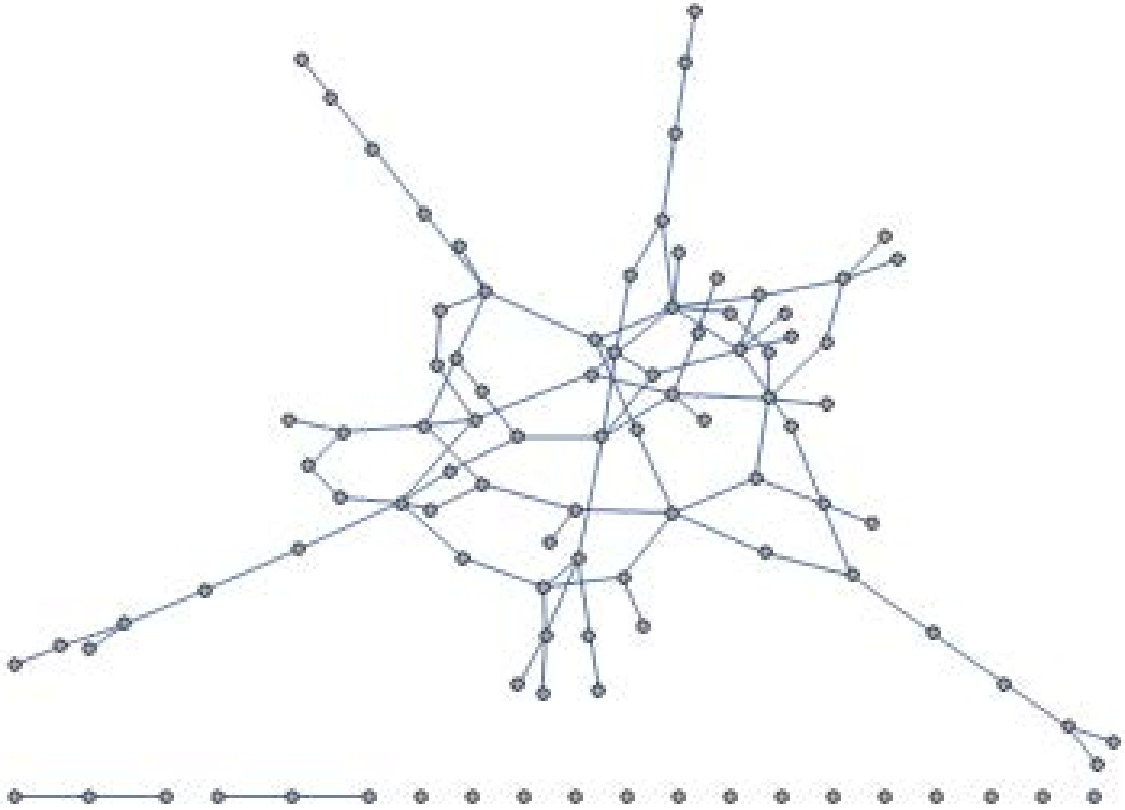
$\mathrm{N} = 100, \langle c \rangle = 1$ – on the phase transition, clusters of size $\mathrm{N}^{2/3}$,

with a substantial number of isolated components



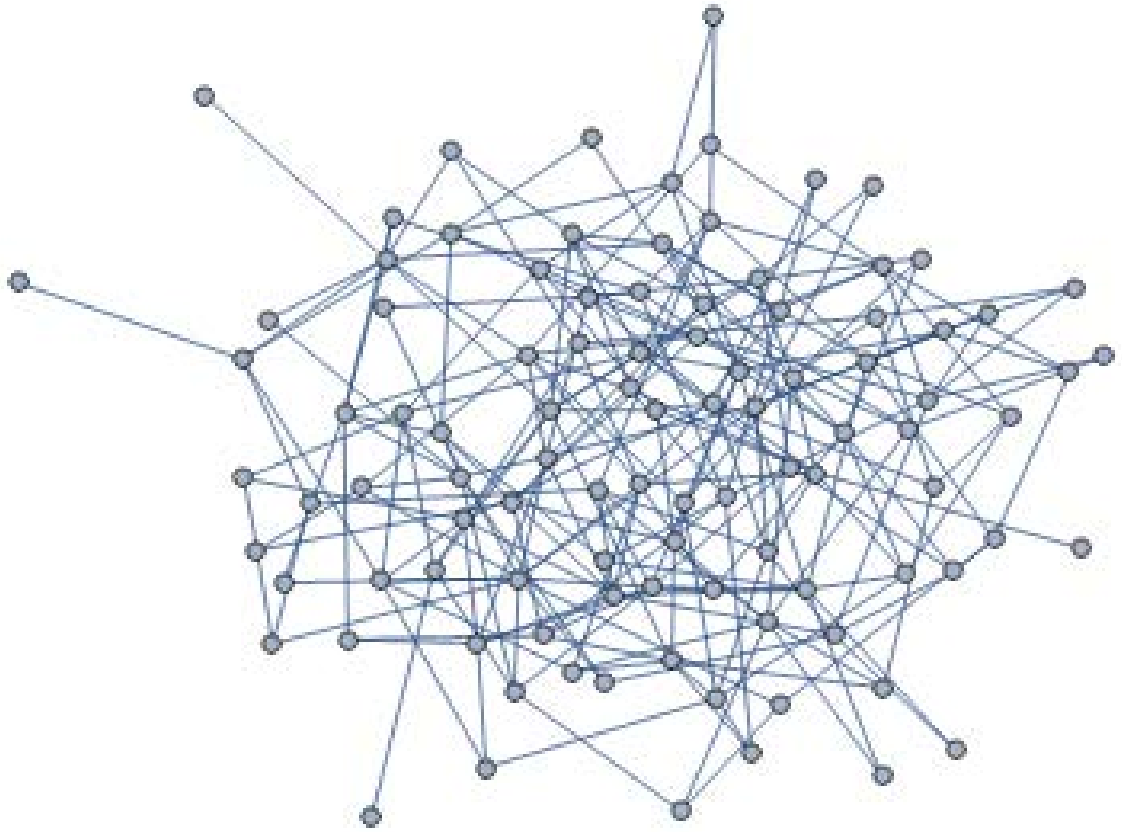
Examples – $\mathrm{ER}(\mathrm{N}, \mathrm{p}),\ \langle c \rangle = 2$

$\mathrm{N} = 100, \langle c \rangle = 2$ – above the phase transition, with a giant cluster of order N, with some isolated components



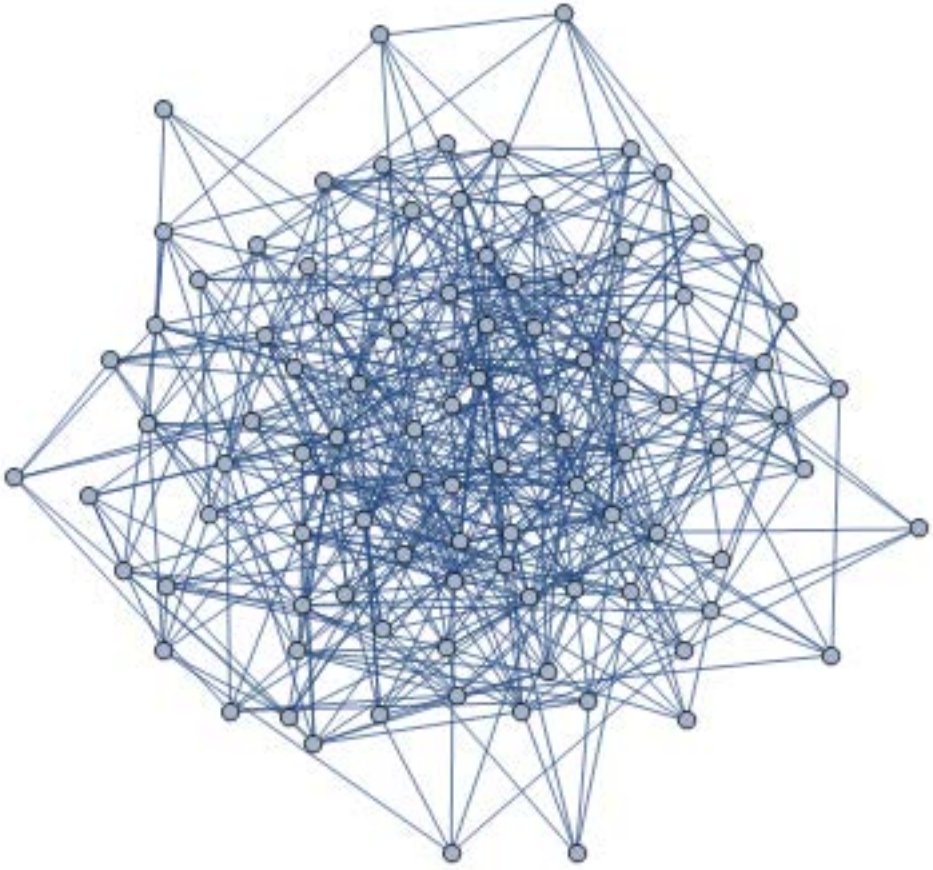
Examples – $\mathrm{ER}(\mathrm{N}, \mathrm{p}),\ \langle c \rangle = 4.6$

$\mathrm{N} = 100, \langle c \rangle = \log N$ – one giant cluster with **no** isolated components. The probability for an isolated node smaller than 1/N



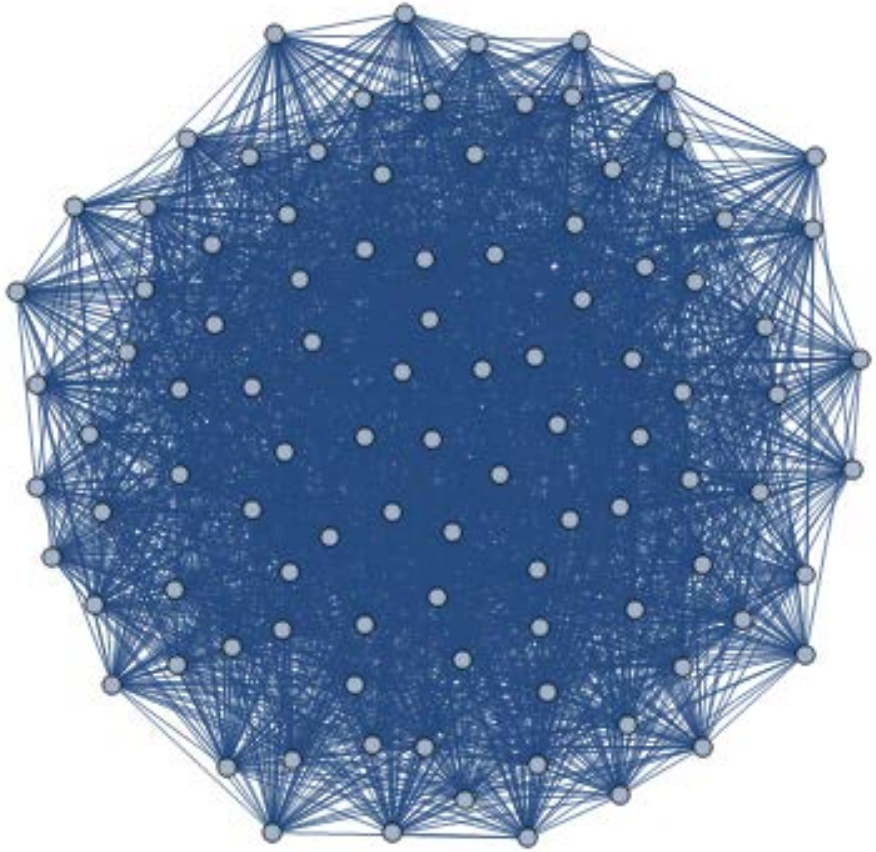
Examples – $\mathrm{ER}(\mathrm{N}, \mathrm{p}),\ \langle c \rangle = 10$

$\mathrm{N} = 100, \langle c \rangle = N^{1/2}$ – dense. One giant cluster with no isolated components.



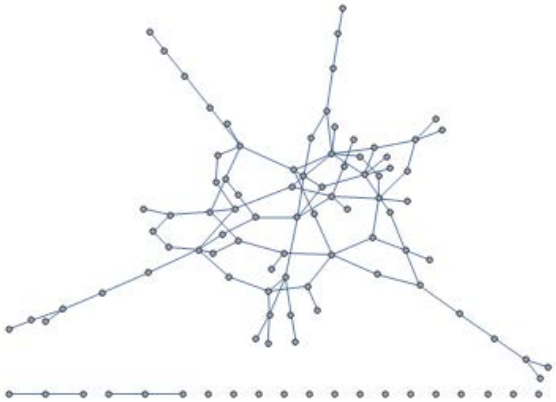
Examples – $\mathrm{ER}(\mathrm{N}, \mathrm{p}),\ \langle c \rangle = 50$

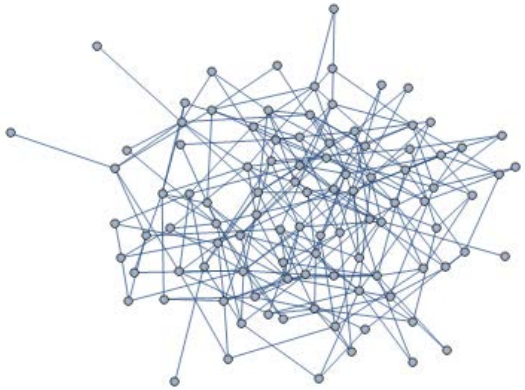
$\mathrm{N} = 100, \langle c \rangle = \frac{1}{2} N$ – very dense.

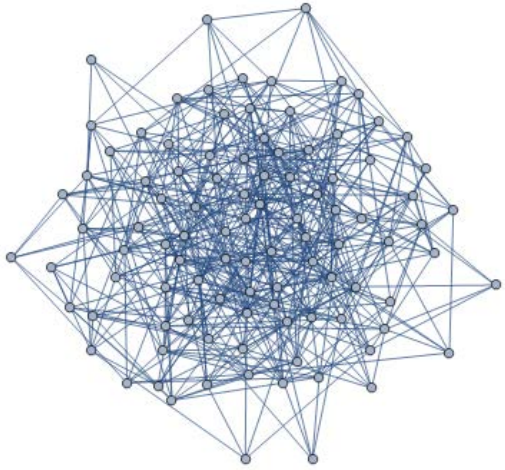


Focus

We will focus in cases where is ‘generally speaking’ a connected network, potentially with some isolated components. The theory collapses below the percolation transition, and works excellently when there are no isolated components at all





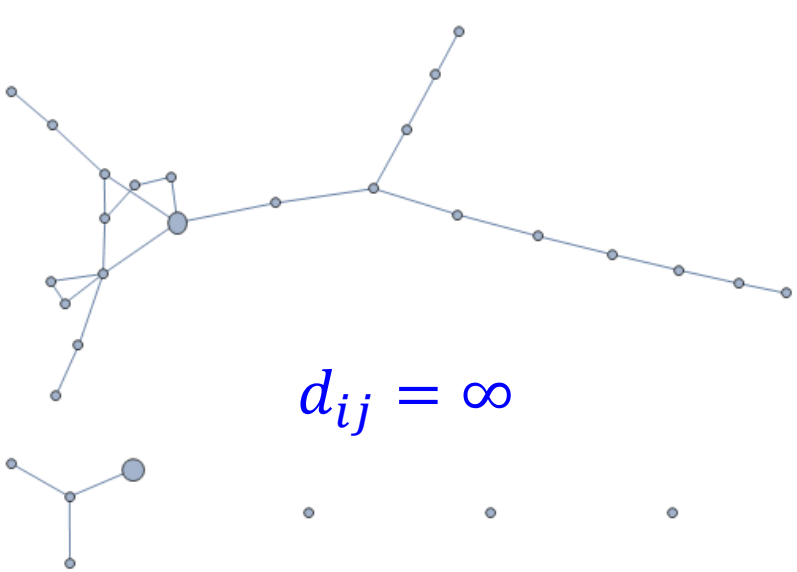


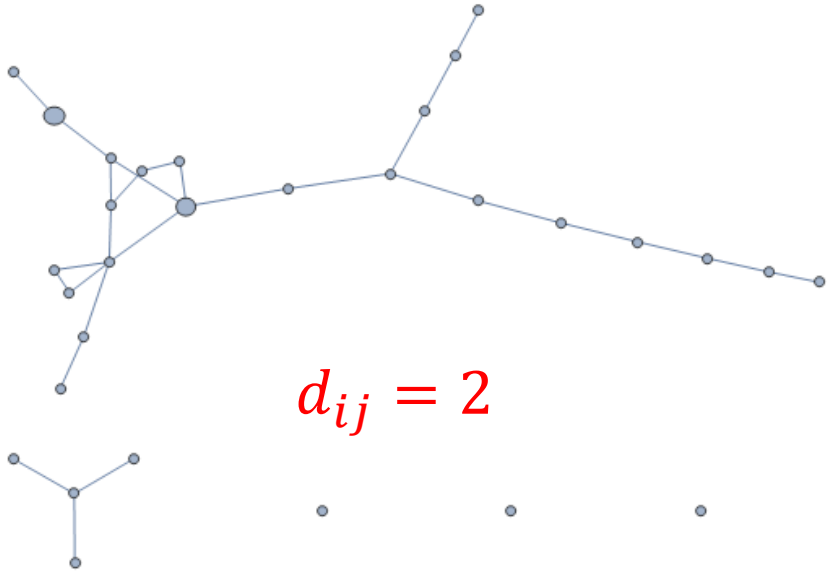
Shortest Paths

Choose two nodes at random i and j, and consider the different paths connecting them.

If these two nodes lie on different clusters then $d_{ij} = \infty$

Otherwise, we focus on the shortest possible path (which needs not be unique) and its length is denoted by d_{ij} (and is finite $d_{ij} < \infty$).
Note that the shortest path on a finite graph cannot exceed $N - 1$.





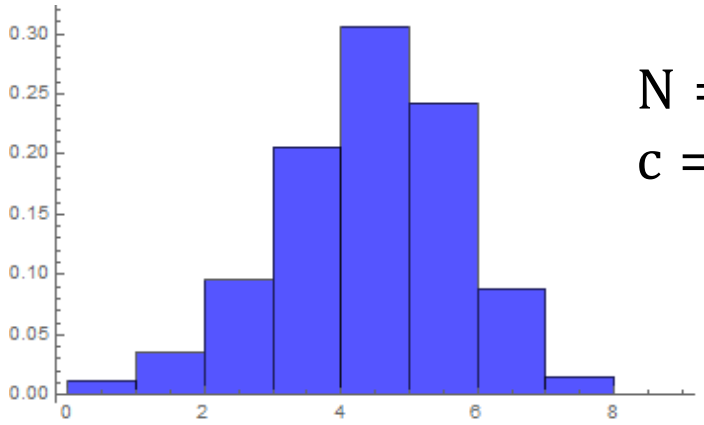
Shortest Paths

The Distribution of Shortest Path Lengths (DSPL) $P_N(d_{ij})$ is simply the distribution of all these distances over the ensemble of random graphs. And we can drop the indices i,j in $P_N(d)$ since it holds for any pair.

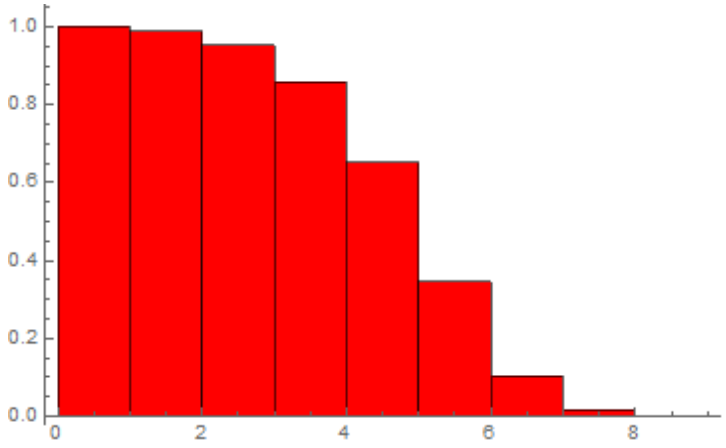
It is sometimes convenient to study the tail-distribution (also known as the survival probability) defined by $F_N(k) = Pr_N(d_{ij} > k)$

e.g. $F_N(1) = Pr_N(d_{ij} > 1) = 1 - p = q$

Last, $P_N(k) = F_N(k-1) - F_N(k)$



$N = 100$
$c = 3$



Shortest Paths – known results

Known results are mainly for the sparse case $\mathrm{p} = \mathrm{c}/\mathrm{N}$

The mean distance is then

$$\langle d \rangle = \sum_{d} d \cdot P_N(d) \simeq \frac{\log N}{\log c} + O(1)$$

which exhibits/explains the small-world phenomenon.

And the diameter, i.e. the longest possible distance among all possible shortest paths in a given network (i.e.),

$$diam\big(ER(N, c/N)\big) = \max\{d_{12}, d_{13}, \ldots, d_{N-1,N}\} \rightarrow \frac{\log N}{\log c} + 1$$

(Bollobás, Janson & Riordan 2007) - essentially the same as the mean!

The full distribution is nevertheless of great interest, but has not been studies even for the simple ER case.

getting started …

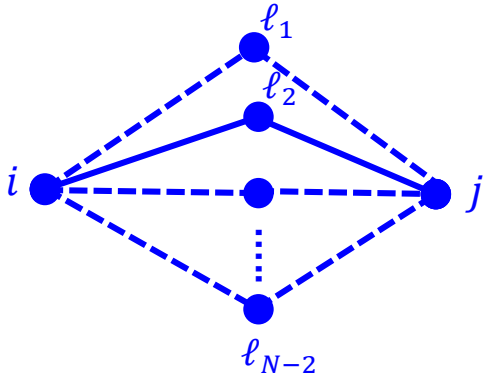
$F_N(1) = P_N(d > 1) = 1 - p = q$

$F_N(2) = P_N(d > 2) = ?$

getting started …

$$F_N(1) = P_N(d > 1) = 1 - p = q$$

$$F_N(2) = P_N(d > 2) = q \times \left(1 - p^2\right)^{N-2}$$ **Exact!**

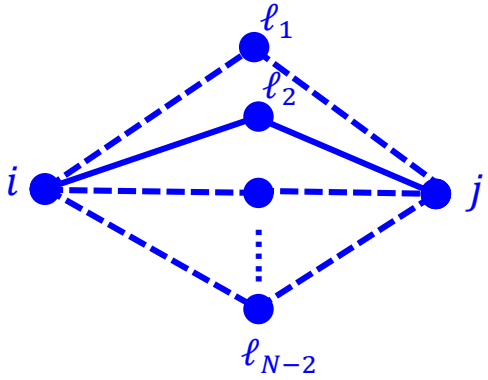


$$F_N(3) = P_N(d > 3) = q \times \left(1 - p^2\right)^{N-2} \times ?$$

getting started …

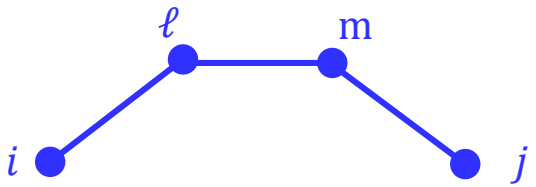
$F_N(1) = P_N(d > 1) = 1 - p = q$

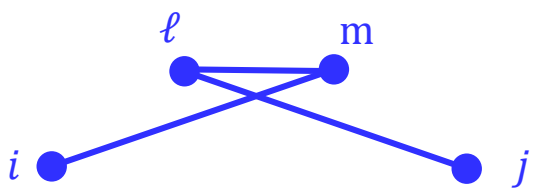
$F_N(2) = P_N(d > 2) = q \times (1 - p^2)^{N-2} = P_N(d > 1) \times P_N(d > 2 | d > 1)$

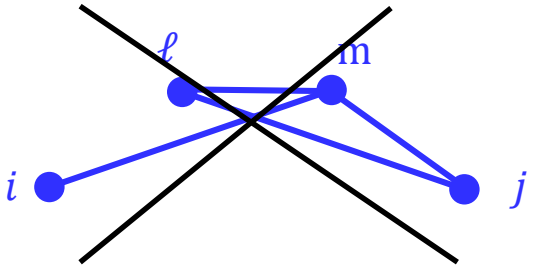


$F_N(3) = P_N(d > 3) = P_N(d > 2) \times P_N(d > 3 | d > 2) = F_N(2) \times P_N(d > 3 | d > 2)$

And the questions is $P_N(d > 3 | d > 2) = ?$



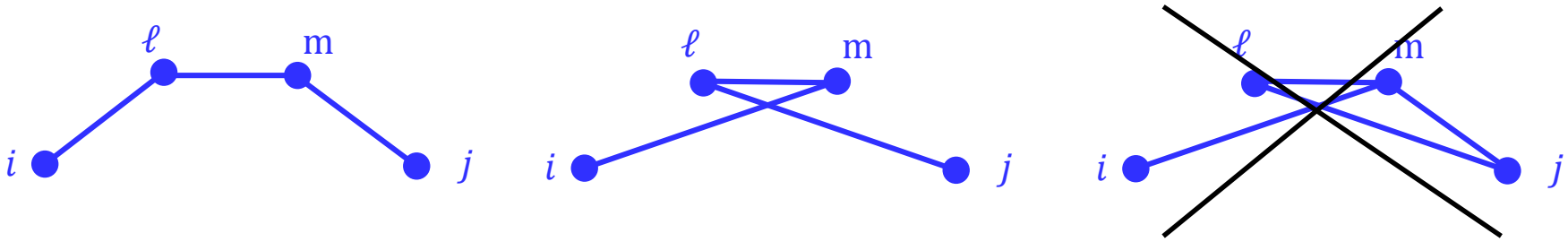




getting started …

$F_N(3) = P_N(d > 3) = P_N(d > 2) \times P_N(d > 3|d > 2) = F_N(2) \times P_N(d > 3|d > 2)$

And the questions is $P_N(d > 3|d > 2) = ?$



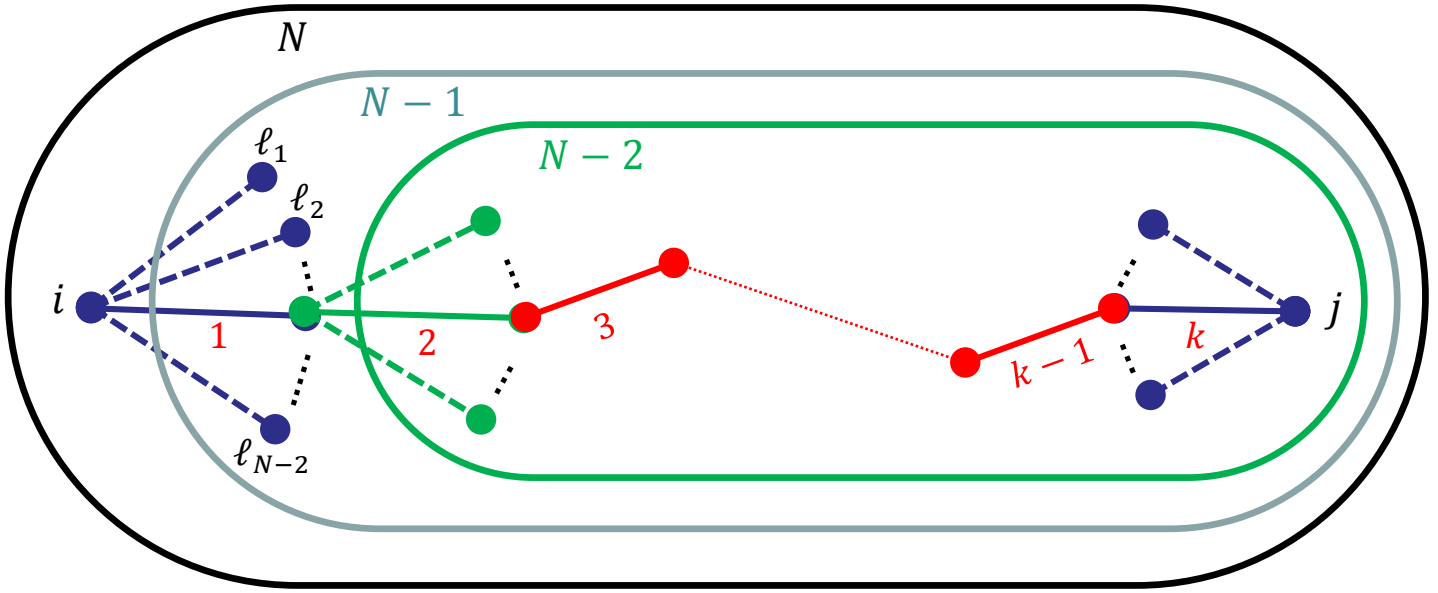
A naïve approach would be $P_N(d > 3|d > 2) \approx (1 - p^3)^{(N-2)(N-3)}$

However this result neglects correlations. Some of these correlations can be handled, but one can convince oneself that the result cannot be made exact for a general N.

$$F_N(k) = F_N(k-1)P_N(d > k|d > k-1) =$$
$$= F_N(k-2)P_N(d > k-1|d > k-2)P_N(d > k|d > k-1) =$$
$$= \cdots = \prod_{m=1}^{k} P_N(d > m|d > m-1)$$

Recursive Paths Approach (RPA)

Let's look at the general case – we can actually relate the conditionals for a system of size N to those of a smaller system of size N-1, but for shorter distances, as follows:



$$P_N(d > k | d > k-1) = [q + pP_{N-1}(d > k-1 | d > k-2)]^{N-2}$$

with $P_{N'}(d > 1 | d > 0) = P_{N'}(d > 1) = q$

Recursive Paths Approach (RPA)

$$P_N(d > k | d > k-1) = [q + pP_{N-1}(d > k-1 | d > k-2)]^{N-2}$$

with $P_{N'}(d > 1 | d > 0) = q$

Comments:

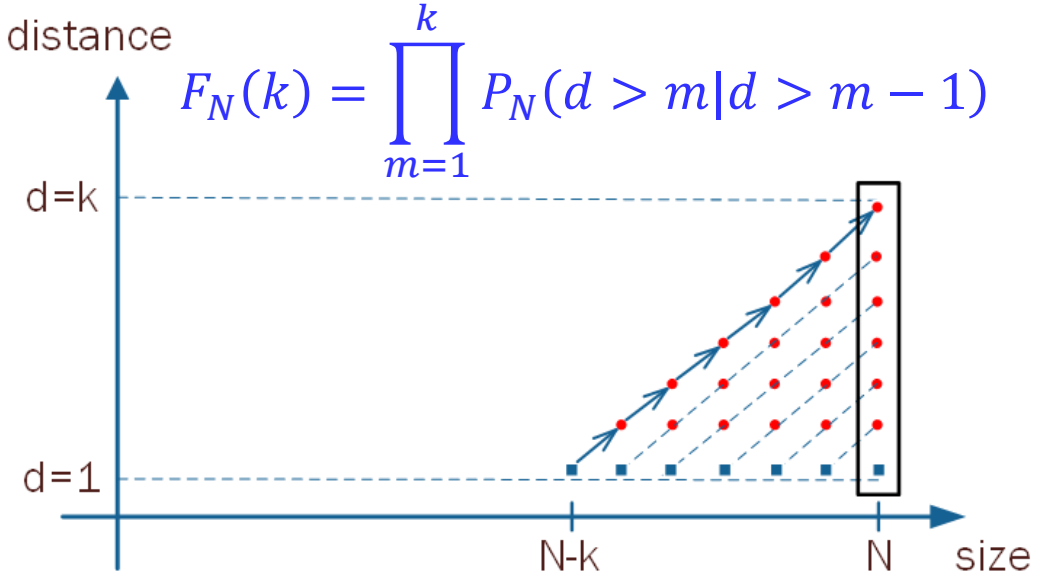
1. It recovers the exact result for $k = 2$, namely

$$P_N(d > 2 | d > 1) = (1 - p^2)^{N-2}$$

2. In the limit of small p: $P_N(d > k | d > k-1) \approx \left(1 - p^k\right)^{(N-2)\ldots(N-k)}$

3. The final expression is obtained using

$$F_N(k) = \prod_{m=1}^{k} P_N(d > m | d > m-1)$$



RPA and the cavity method

The RPA shares the same spirit with the cavity method. This can actually be made more formally, and help further progress.

The jPDF of the ER network: $P(C) = \prod_{i<j} \left[(1-p)\delta_{c_{ij},0} + p\delta_{c_{ij},1}\delta_{c_{ji},1}\right]$

And the starting point are the indicator functions $\chi\left(d_{ij} > k | d_{ij} > k-1\right)$

One can write down the recursion equations for these quantities

$$\chi\left(d_{ij} > k | d_{ij} > k-1\right) = \prod_{\ell \neq i,j} \left[1 - c_{i\ell} + c_{i\ell}\chi^{(i)}\left(d_{\ell j} > k-1 | d_{\ell j} > k-2\right)\right]$$

where $\chi^{(i)}\left(d_{\ell j} > k-1 | d_{\ell j} > k-2\right)$ are the cavity indicator functions obeying a similar equation.

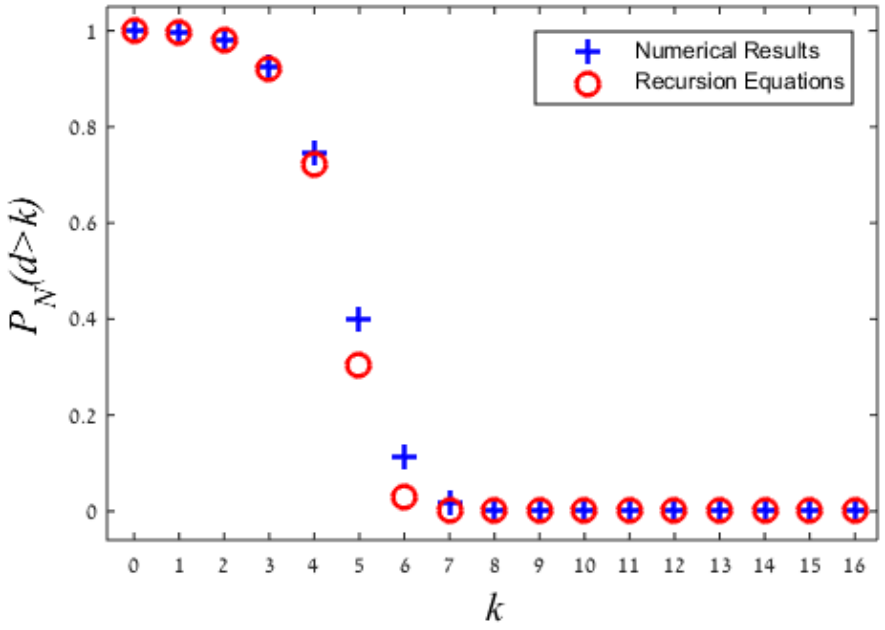
One can average over these equations using $P(C)$ and obtain the RPA equations for the Erdos-Renyi case.

As usual, the regular graph case is straightforward.

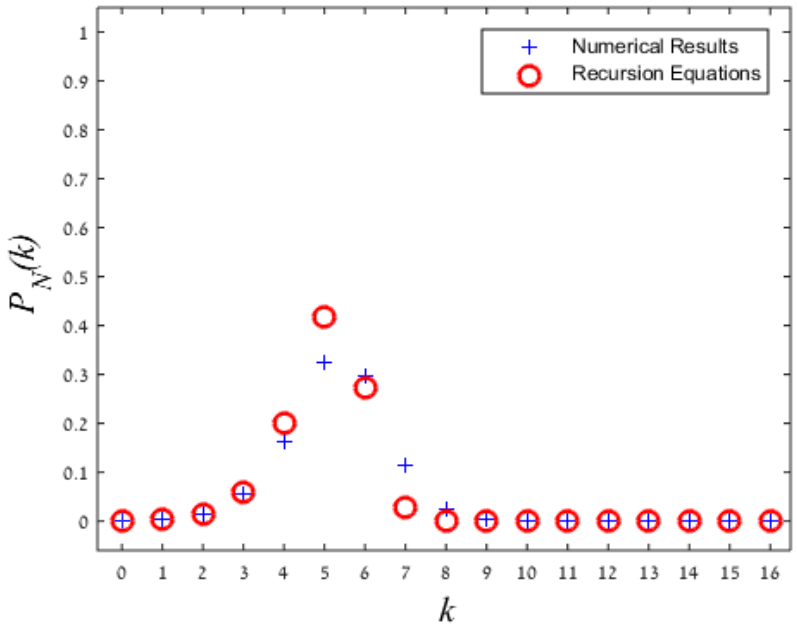
RPA – results

$$\mathrm{N} = 1000, \mathrm{c} = 4 \; - ER\left(1000, \frac{4}{1000}\right)$$

The tail distribution



The PDF

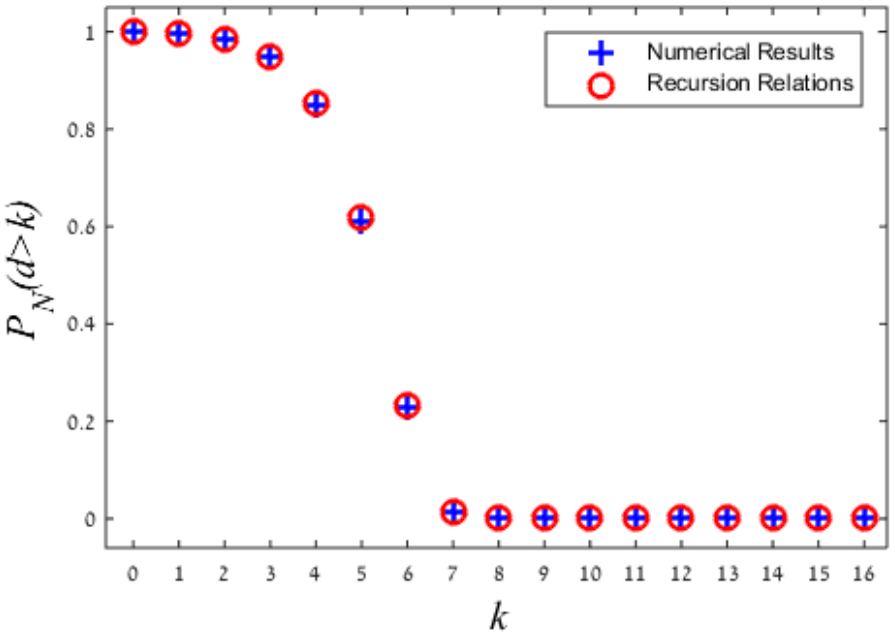


The agreement between the numerical result and the RPA equation is quite good already for $\mathrm{N} = 1000$.

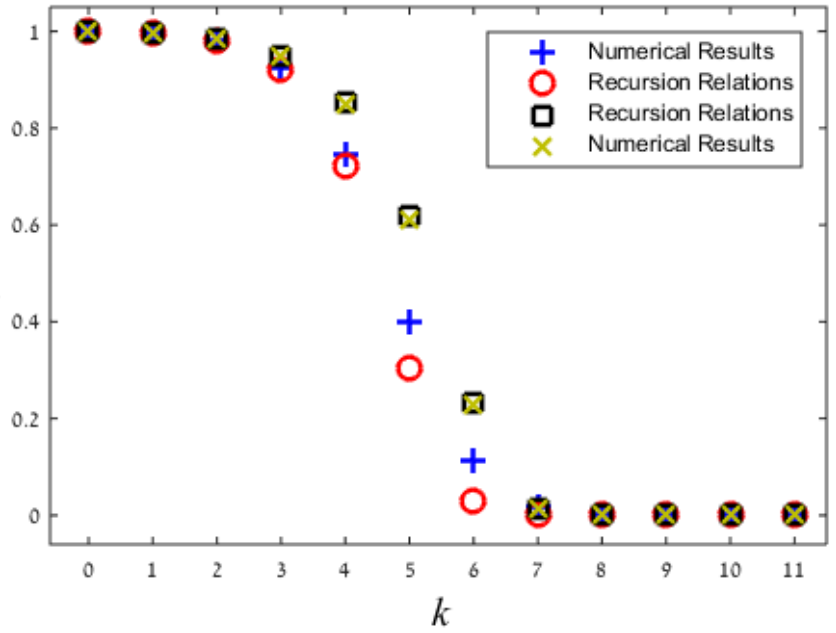
RPA – results

$N = 1000, c = 4$, i.e. 4-regular graph

Regular Graph



Erdos-Renyi + Regular



The agreement in the case of a regular graph is much better.

Also note that distances in the ER graph are slightly shorter than in a regular graph with the same mean degree.

RPA – the dense case $c = bN^{\alpha}$

In this case the network is dense but p is still vanishingly small.

The prediction of the theory is that there is a categorical difference between $\alpha = 1/r$ where $r \in Z^{+}$ (i.e. $r = 2,3,4 \ldots$) and other values of α namely:

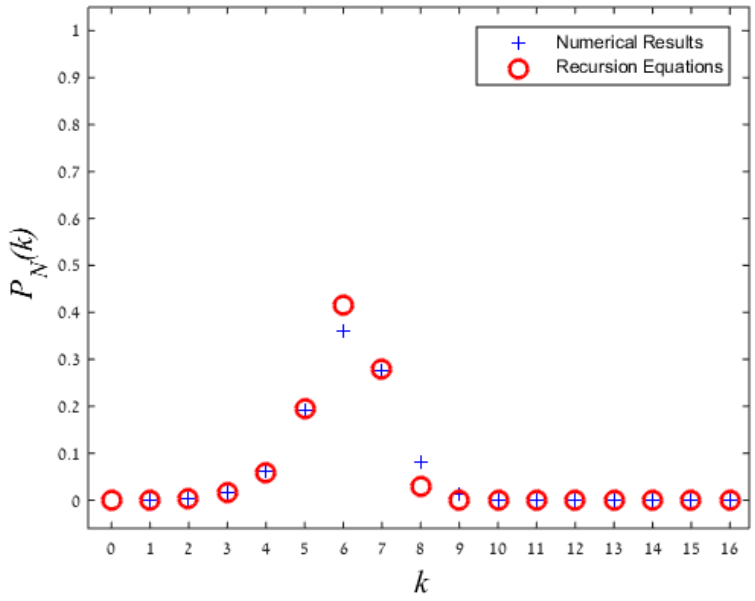
$$P_{N\to\infty}(k) = \begin{cases} \begin{cases} 1 & for\ k = \left\lfloor \frac{1}{\alpha} \right\rfloor + 1 \\ 0 & otherwise \end{cases} & if\ \alpha \notin \left\{\frac{1}{2}, \frac{1}{3}, \ldots\right\} \\ \begin{cases} 1 - e^{-b^{1/\alpha}} & for\ k = \frac{1}{\alpha} \\ e^{-b^{1/\alpha}} & for\ k = \frac{1}{\alpha} + 1 \\ 0 & otherwise \end{cases} & if\ \alpha \in \left\{\frac{1}{2}, \frac{1}{3}, \ldots\right\} \end{cases}$$

The distribution becomes extremely narrow. Generically there is only one non-vanishing entry, apart from the special case $\alpha = 1/r$ where there are two such values.

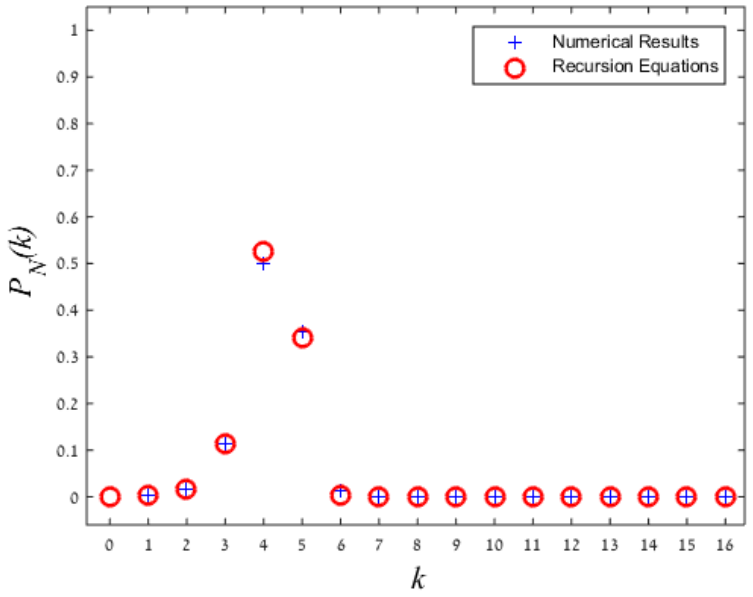
RPA – the dense case $c = bN^{\alpha}$

A new prediction - almost all pairs are at the same distance from each other. This is the closest one can get to a regular-distance graph. (One needs large networks to realize the asymptotic limit)

$N = 4096, c = 16, \alpha = 1/6$

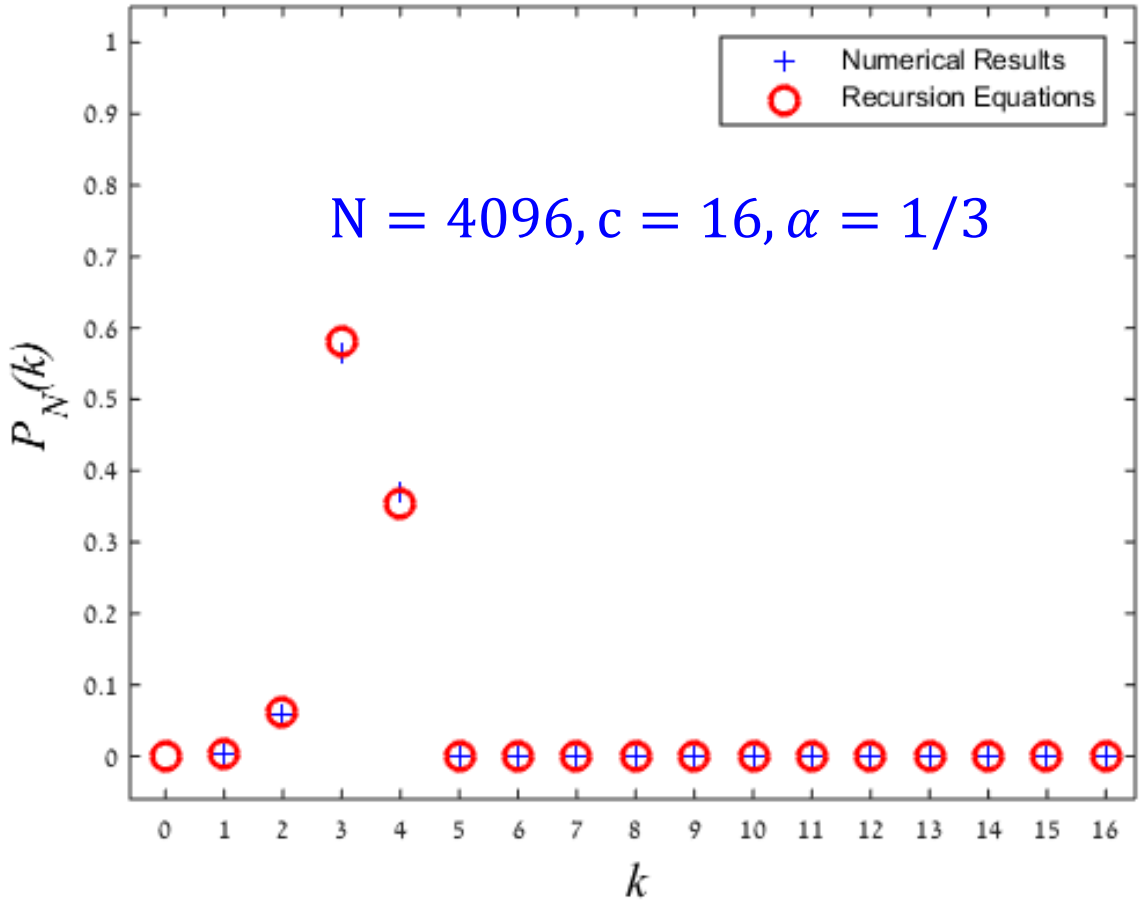


$N = 4096, c = 8, \alpha = 1/4$

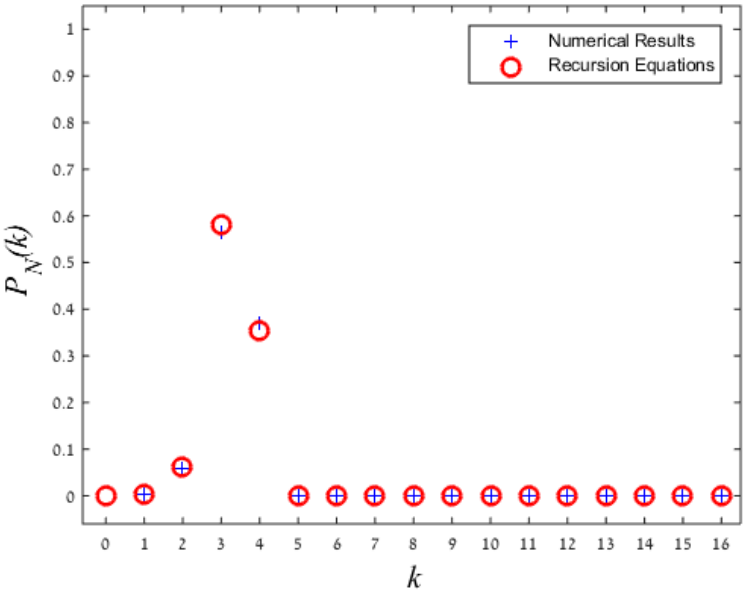


RPA – the dense case $c = bN^{\alpha}$

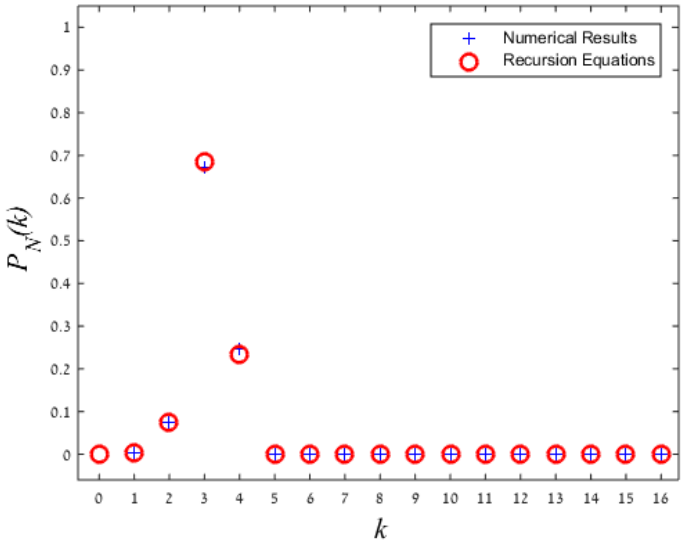
A new prediction - almost all pairs are at the same distance from each other. This is the closest one can get to a regular-distance graph.



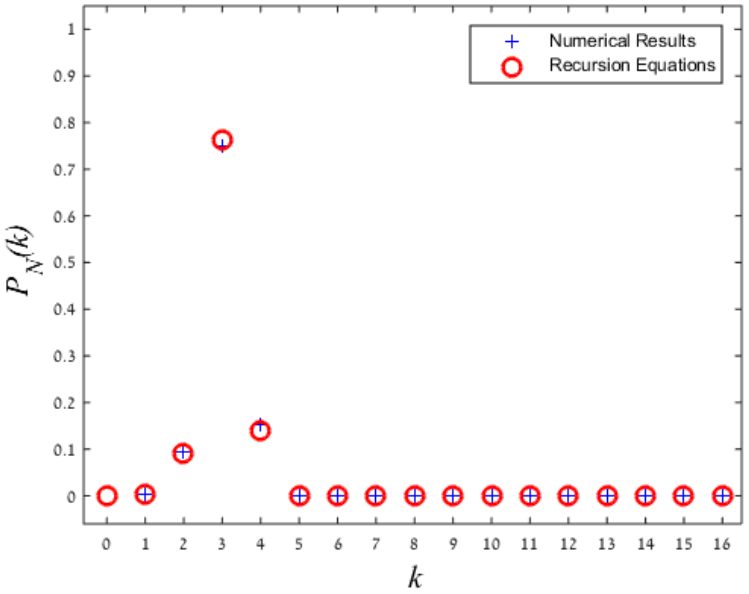
$$\mathrm{N} = 4096, \mathrm{c} = 16, \boldsymbol{\alpha = 0.3333}$$



$$\mathrm{N} = 4096, \mathrm{c} = 18, \boldsymbol{\alpha = 0.347}$$



$$\mathrm{N} = 4096, \mathrm{c} = 20, \boldsymbol{\alpha = 0.36}$$



Summary

We studied the distribution of shortest path lengths/distances in ER networks using a recursive path understanding of the network.

The numerical results agree well with the theory above the percolation threshold $c = 1$, but well still when there are isolated components. It becomes really excellent in the dense limit.

Distances are typically (very) short, and the distribution is narrow.

The results for $P_N(d > 1)$, $P_N(d > 2)$ are exact.

Can be extended to other random graph ensembles (e.g. the configuration mode), maybe improved, and can have many applications – e.g. in dynamical processes on networks, inference and sampling of networks with desired distance distribution.

We have an analytical handle that can serve further riddles.