The distribution of shortest path lengths in random networks

Eytan Katzav

The Racah Institute of Physics

The Hebrew University of Jerusalem

In collaboration with **Reimer** and with
Ofer Biham, Ido Tishby, Chanania Steinbock and Barak Budnick

Disordered Systems Days at King's College London, 11-12 Sep 2023

Outline

The **D**istribution of **S**hortest **P**ath **L**engths – **DSPL** in random networks

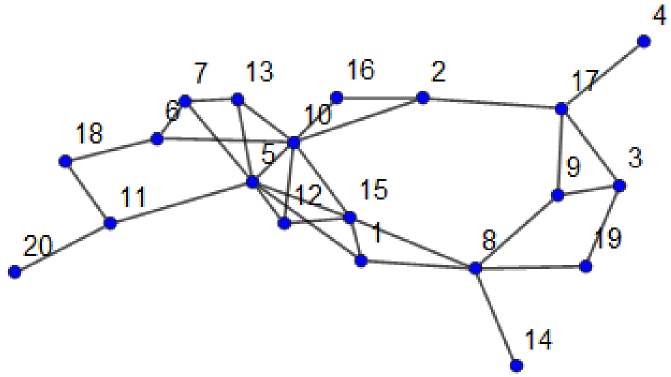
- Random Graphs: Equilibrium (Erdős–Rényi, Regular, Configuration model etc..) and out-of-equilibrium networks (BA and Node-Duplication)
- Metric properties of random graphs:

  Shortest Paths – definition, known results
- The importance of the DSPL in dynamical processes
- Review of the analytical approaches we developed:

  1. Equilibrium networks: The Recursive Paths Approach (RPA) – Cavity / Message-Passing or Renormalization
  2. Cycles, and exact results for sub-critical ER.
  3. Out of equilibrium networks: Master equation for node duplication.

Graphs/Networks

Graphs are essentially a collection of vertices (nodes) and edges (links). One could think of weighted graphs, directed graphs and more

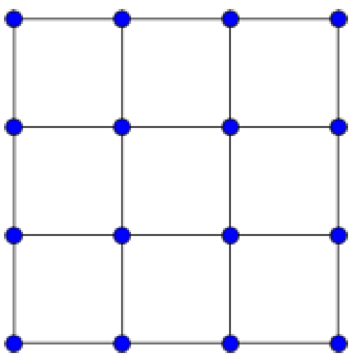


$$V = \{1,2,\ldots,20\}$$

$$E \subseteq V \times V$$

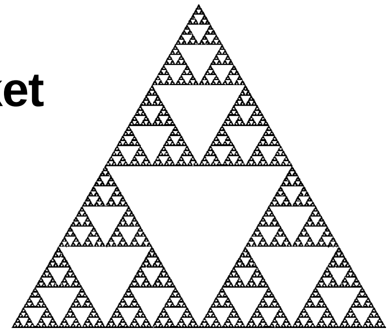
$$G = (V, E)$$

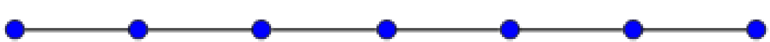
Regular lattices as well as fractals are also graphs, and actually very special ones



Sierpinski Gasket

$d_F = 1.58\ldots$

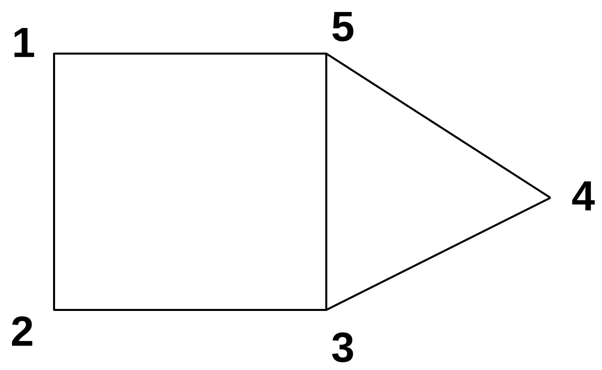




Graphs and Matrices

Graphs and Matrices are intimately related, as there are a few matrix representations of graphs, which provide powerful tools to analyse them:

1. Adjacency Matrix



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix} \leftrightarrow A_{ij} = \begin{cases} 1 & \text{nodes i and j are linked} \\ 0 & \text{otherwise} \end{cases}$$

2. Graph Laplacian

$$L = I - D^{-1/2} A D^{-1/2} \quad \left(D = \text{degree matrix} = \text{diag}\{d_1, \ldots, d_N\} = \text{diag}\{2,2,3,2,3\} \right)$$

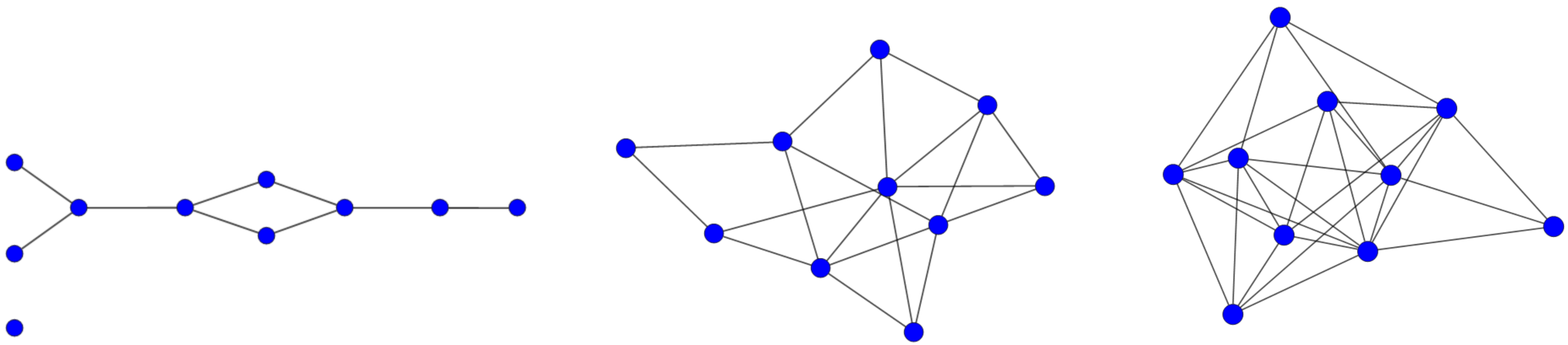
3. Modularity Matrix

$$M_{ij} = A_{ij} - \frac{d_i d_j}{2m} \qquad \left(m = \sum_i d_i \right)$$

(these matrices are typically sparse)

Random Graphs/Networks

There are many interesting families of random graphs, examples being – Erdős–Rényi $\mathrm{ER}(N, p)$ graphs (N nodes, and probability p for the existence of every possible edge independently), e.g. for $N = 10$

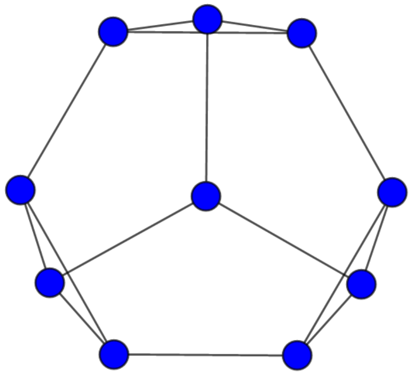


$p = 0.2$ $\quad$ $p = 0.4$ $\quad$ $p = 0.6$

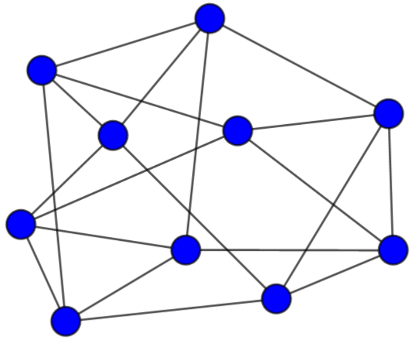
Binomial (Poisson) degree distribution $P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$

Random Regular Graphs

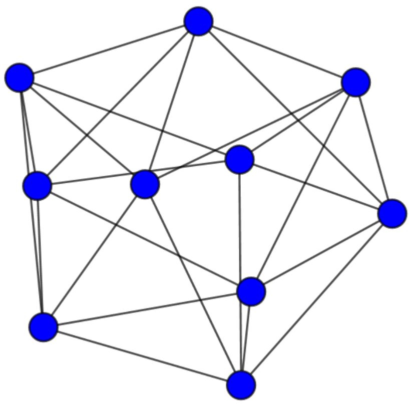
Random regular graphs (N nodes, all vertices of degree c, and a uniform sampling out this ensemble), e.g. $N = 10$, $P(k) = \delta_{k,c}$



$c = 3$
(cubic)



$c = 4$



$c = 5$

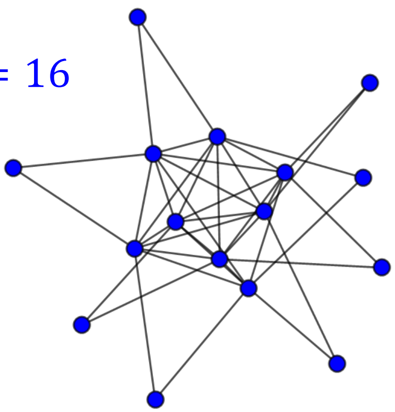
Such graphs often enjoy many analytical results.

They also share properties in common with lattices in low dimensions, e.g. a low coordination number and regularity.

The Configuration Model

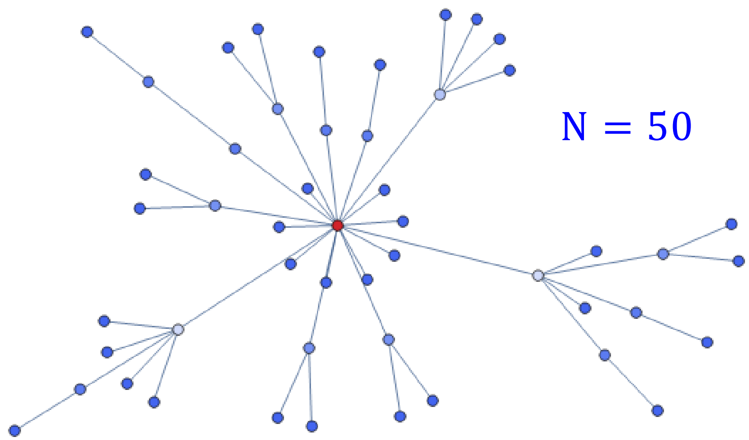
More generally one can impose any (admissible) degree sequence, e.g. given by a certain **degree distribution $P(k)$**. This can be any distribution such as exponential, Gaussian or even scale-free, but with no further correlations. Technically this is a *maximum entropy ensemble*, and is referred to as the **Configuration model**.

N = 16



$$P(k) = \frac{1}{2}\left(\delta_{k,2} + \delta_{k,8}\right)$$

N = 50



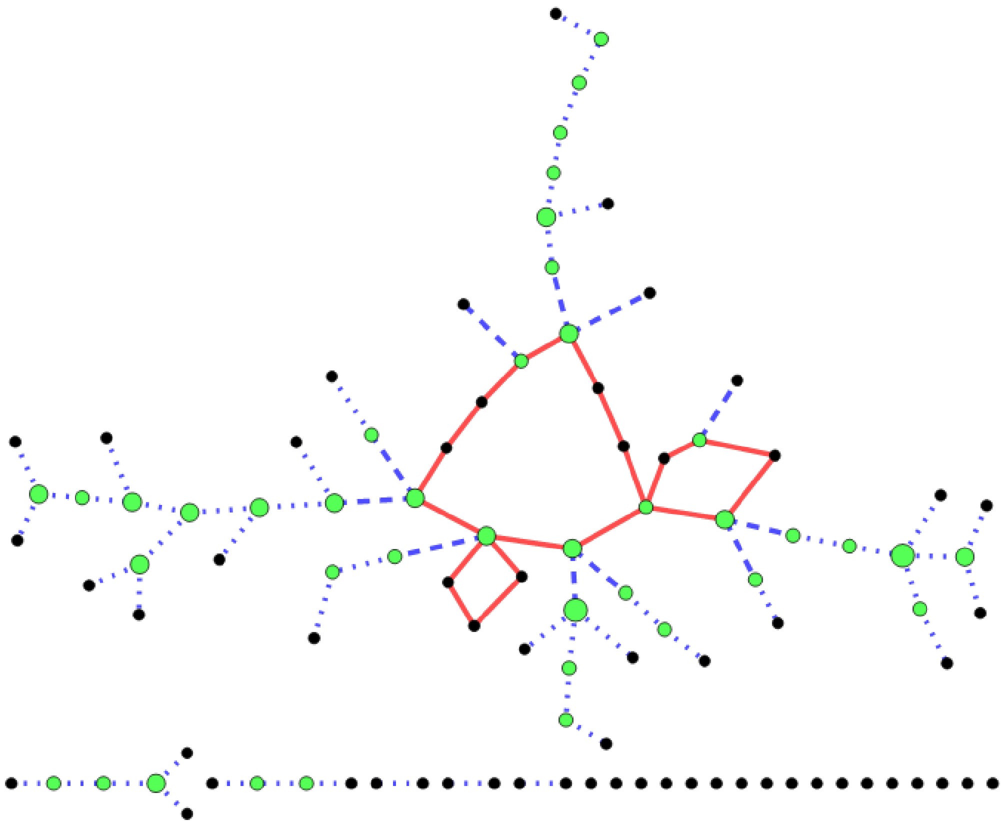
$$P(k) = A\,k^{-\gamma}$$

scale-free

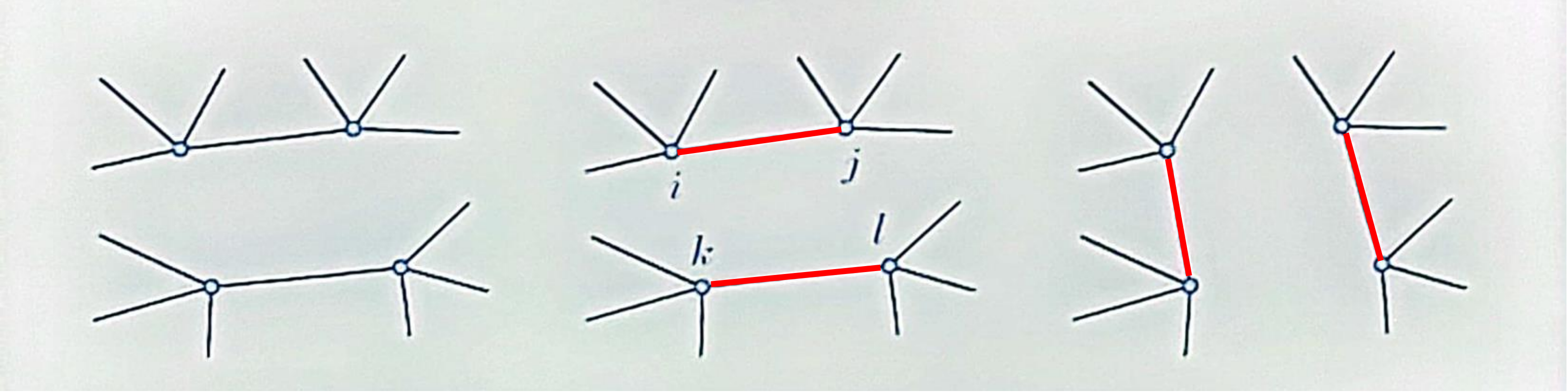
Our approach can be applied to this model.

The Configuration Model

The following example is typical of configuration model networks. It demonstrates common geometrical features, such as the distinction between a giant cluster, and finite components, as well as concepts like the 2-core, Articulation Points (APs), Bredges (bridge-edges) and more



Equilibration: switching



Note that switching does not alter the degree distribution!

Configuration network ensembles are invariant with respect to such transformations. In fact such steps are “equilibration” transformations, in the sense that if you start with any network it converges onto the uncorrelated configuration network ensemble.

Maximum entropy ensemble

Formally, we can define the configuration model as an ensemble of graphs with a probability measure over all possible simple graphs $P(G)$, such that

$$\sum P(G) = 1,$$

and also obeys the constraint that the degree distribution is fixed

$$\sum p(K=k)P(G) = \langle p(K=k)\rangle.$$

The measure can then be obtained by maximizing over the Gibbs entropy

$$S = -\sum P(G)\ln P(G).$$

Actually, in the configuration model this entropy could be simply understood as the (log of the) number of possible graphs with the given degree distribution [Coolen *et al.* 2009]

Kulback-Leibler divergence or relative entropy

$$S = S^{ER} \underline{- \sum p(K=k)\ln\frac{p(K=k)}{p^{ER}(K=k)}}$$

Equilibrium vs. Nonequilibrium

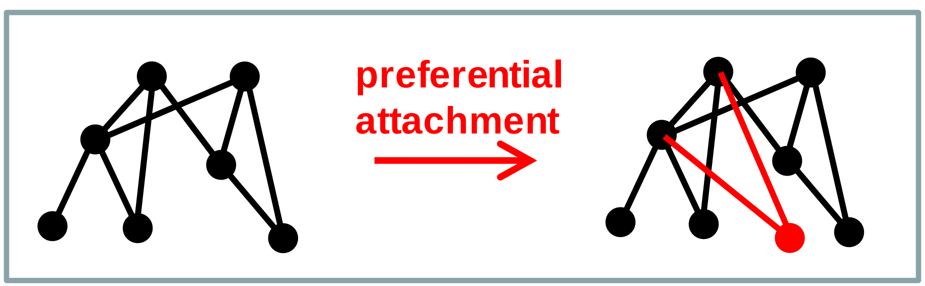
Up to now we have seen only <u>equilibrium</u> ensembles of random networks - just like in the classical ensembles of statistical mechanics they are <u>maximum entropy ensembles</u>, i.e. maximally random on top of some imposed constraints (usually related to degrees).

In contrast, one can design <u>nonequilibrium</u> ensembles, such as <u>growing networks</u> out of equilibrium.

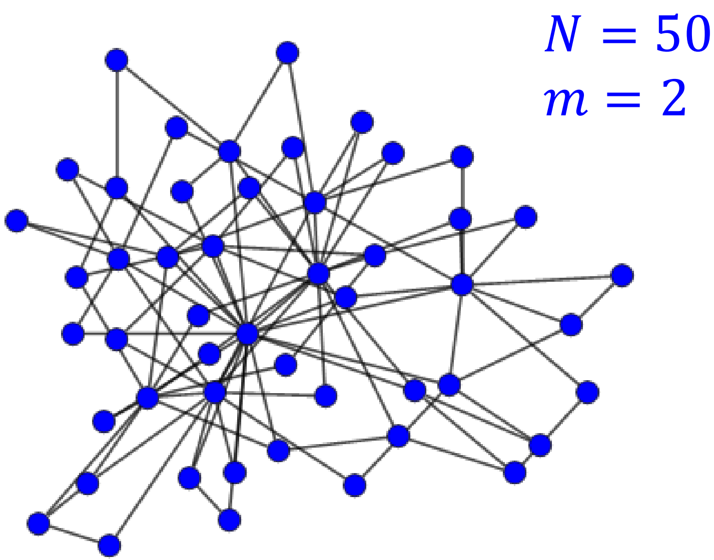
Two prominent examples are **Preferential Attachment** (or the Barabási-Albert model) and the **Node-Duplication** model.

Preferential Attachment

At each time step a new node is added to the network and is randomly connected to a fixed number, $\boldsymbol{m}$, of existing nodes. the probability of each existing node to be connected to the new node is linearly proportional to its degree. The preferential attachment model, **AKA Barabási-Albert model**, gives rise to a power law degree distribution of the form $P(k) = A\,k^{-\gamma}$**,** with $\gamma = 3$**,** which endows the resulting network with a scale-free structure, with a bounded mean and a diverging variance.



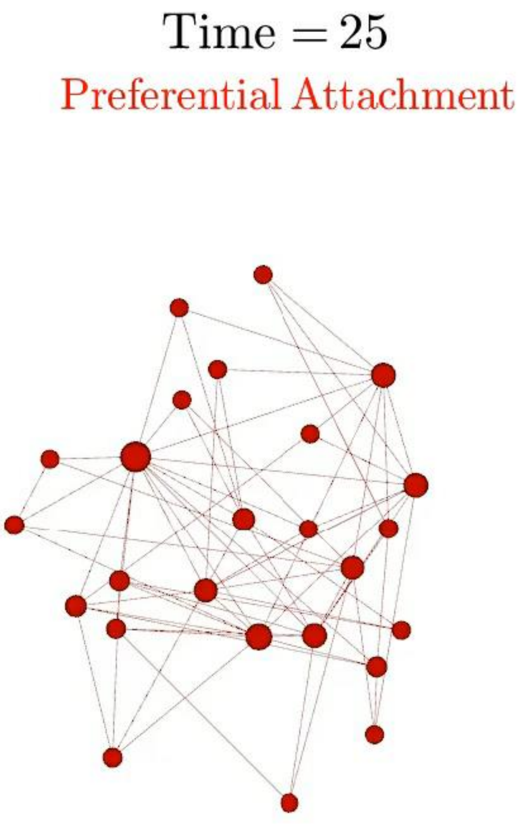
$$P(k) = A\,k^{-3}$$



Preferential Attachment

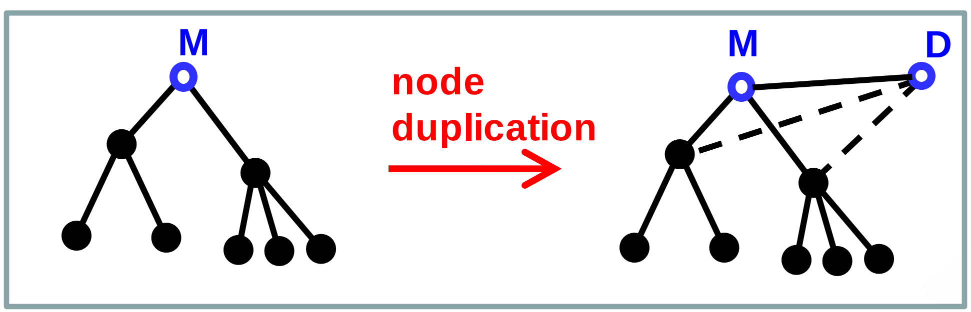
1. By construction it is composed of a single cluster.
2. The rich get richer.

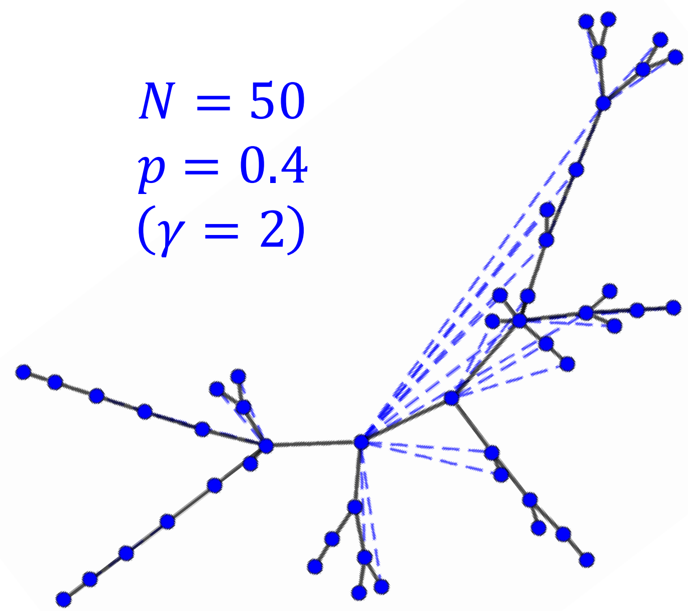
$$P(k) = A\, k^{-3}$$



Node Duplication

Node duplication models are of great importance for the study of social networks as well as **citation networks** in scientific publications. In these models, each time step a random node is duplicated and each one of its links is cloned with probability p. The resulting network exhibits a power-law degree distribution $P(k) = A\,k^{-\gamma}$, in which the exponent γ depends on p.





$$P(k) = A\,k^{-\gamma} \qquad \left(\gamma = 3 - \frac{W(p \ln p)}{\ln p}\right)$$

Why are they interesting?

Insight into generic properties of certain networks, such as the small-world property of random scale-free networks:
explaining the essence of the **'six degrees of separation'** phenomenon (everyone and everything is six or fewer steps away from each other).

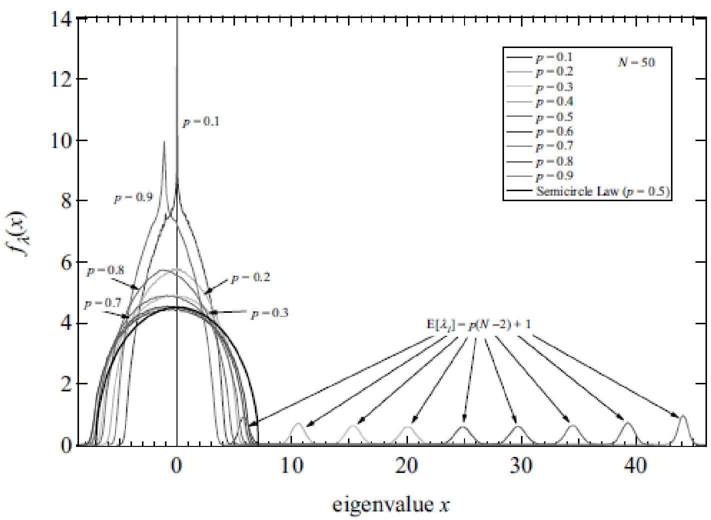
Statistical inference - **null models for real networks**. Suppose you have a real network with a given degree distribution, and you observe a certain pattern. A meaningful way to test for the robustness of this observation would be to compare it to a random null-model or a benchmark.

Can give **insight into low dimensional glassy / disordered systems** using mean-field methods (e.g., cavity or replicas) in the sparse limit.
In many ways random networks resemble high dimensional spaces.

Network geometry – going beyond Euclidean lattices and fractals.
Why do we live in a 3D (4D, 10D…) space?

Relation to random matrix theory

Random networks are intimately related to random matrix theory via the different matrices (Adjacency, Laplacian, Modularity …). In particular spectral properties are shared thanks to universality:



The spectral density of the ER ensemble approaches the Wigner semi-circle law, apart from the largest one:

$$\langle \lambda_{\max} \rangle = (N-2)p + 1 + O\left(1/\sqrt{N}\right)$$

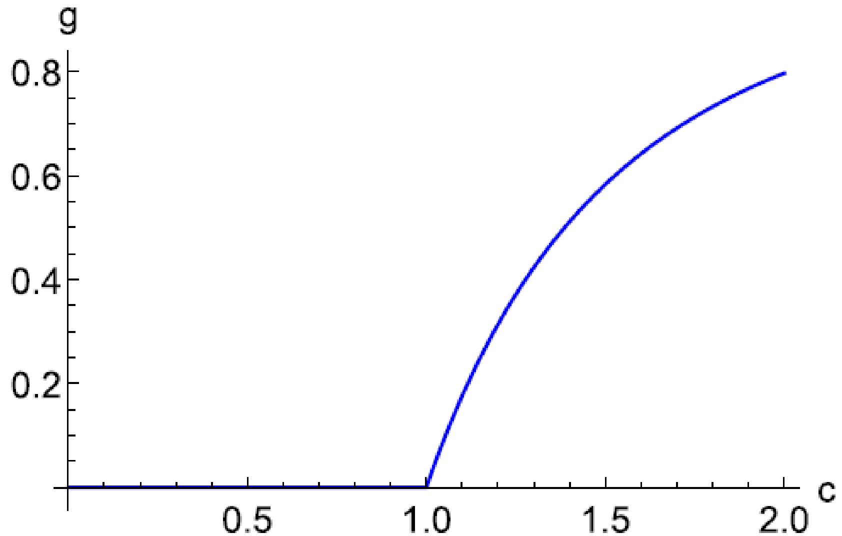
$$\rho(\lambda) \simeq \frac{\sqrt{4Np(1-p)-(\lambda+p)^2}}{2\pi Np(1-p)} \mathbb{I}_{|\lambda| \leq 2p(1-p)\sqrt{N}}$$

Various powerful conclusions follow – such as distribution of first return times, number of triangles, PageRank centrality measure, community detection and more…

Percolation Transition

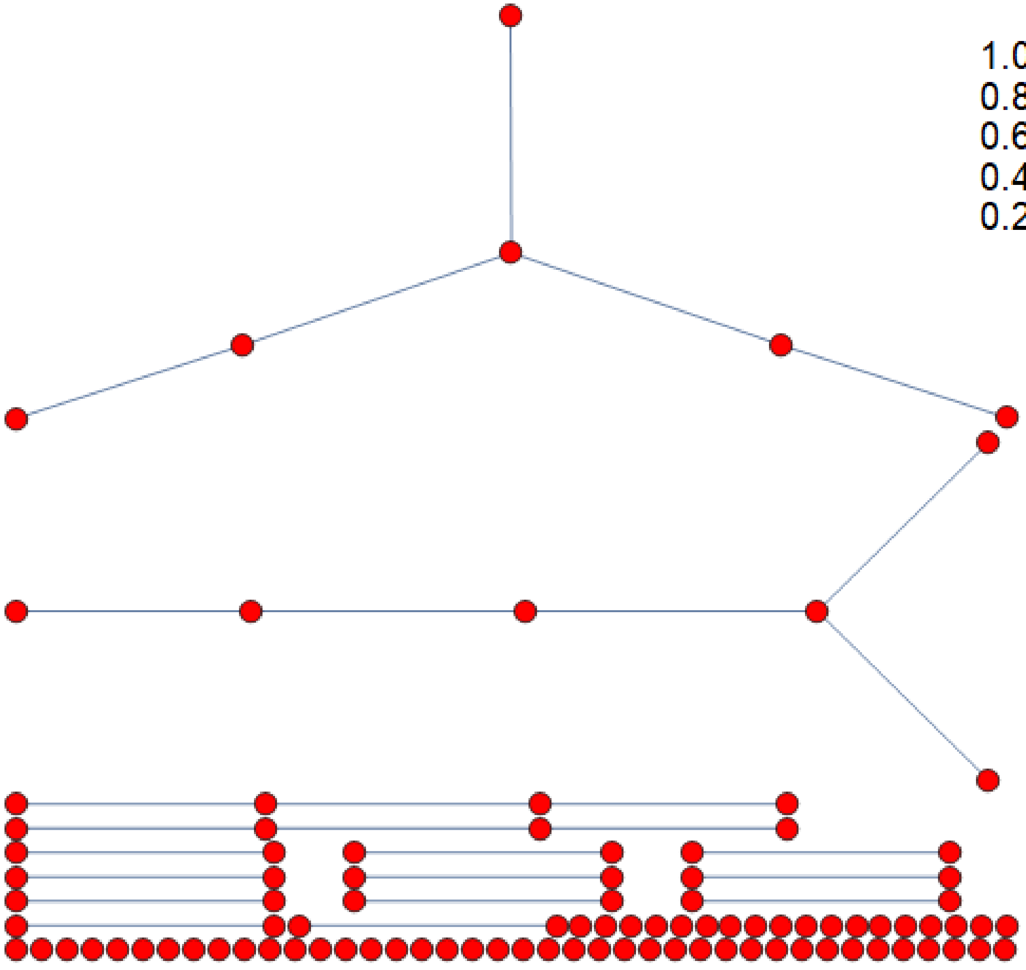
For the sake of simplicity we will focus on ER networks.

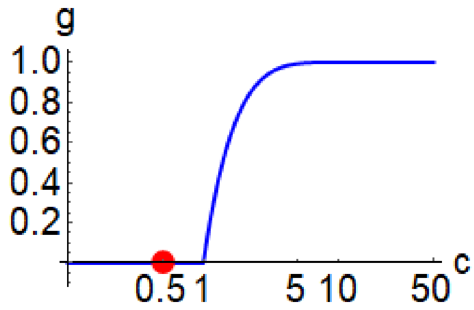
In the following slides I will present a series of ER networks, with a given fixed size ($\mathrm{N} = 100$), and increasing mean value of neighbours $\mathbf{c} = \langle \boldsymbol{K} \rangle = \boldsymbol{Np}$ (or vice-versa, i.e. choose $p = c/N$), starting with $\mathrm{c} = 0.5$ and up to $\mathrm{c} = 50$, in order to gain some acquaintance



Examples – $\mathrm{ER}(\mathrm{N}, \mathrm{p}),\ c = 0.5$

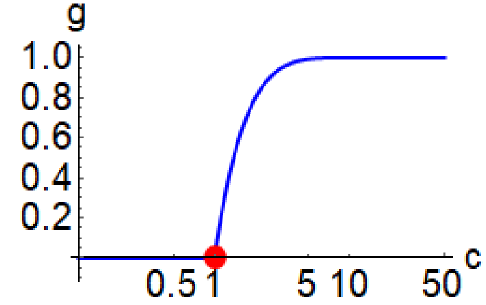
$\mathrm{N} = 100, c = 0.5$ – mostly isolated tree components, a forest

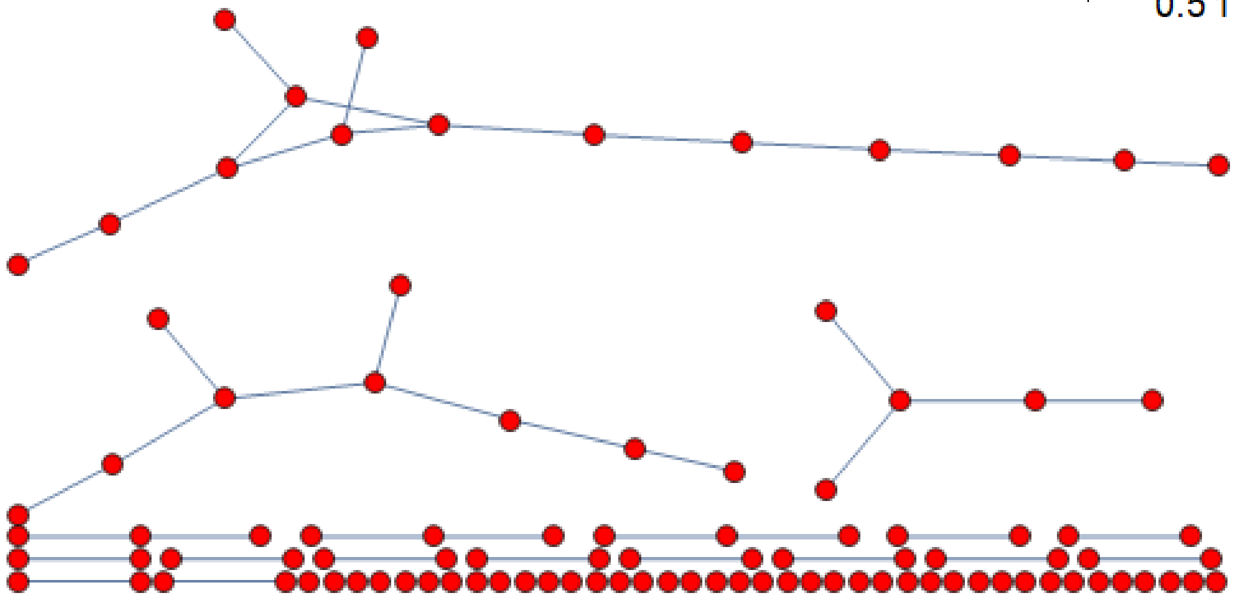




Examples – $\mathrm{ER(N,p)},\ c = 1$

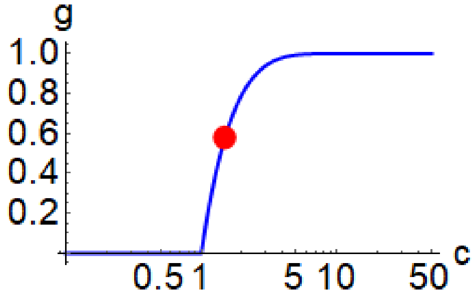
$\mathrm{N} = 100, c = 1$ – on the phase percolation transition, clusters of size $\mathrm{N}^{2/3}$, with a substantial number of isolated components

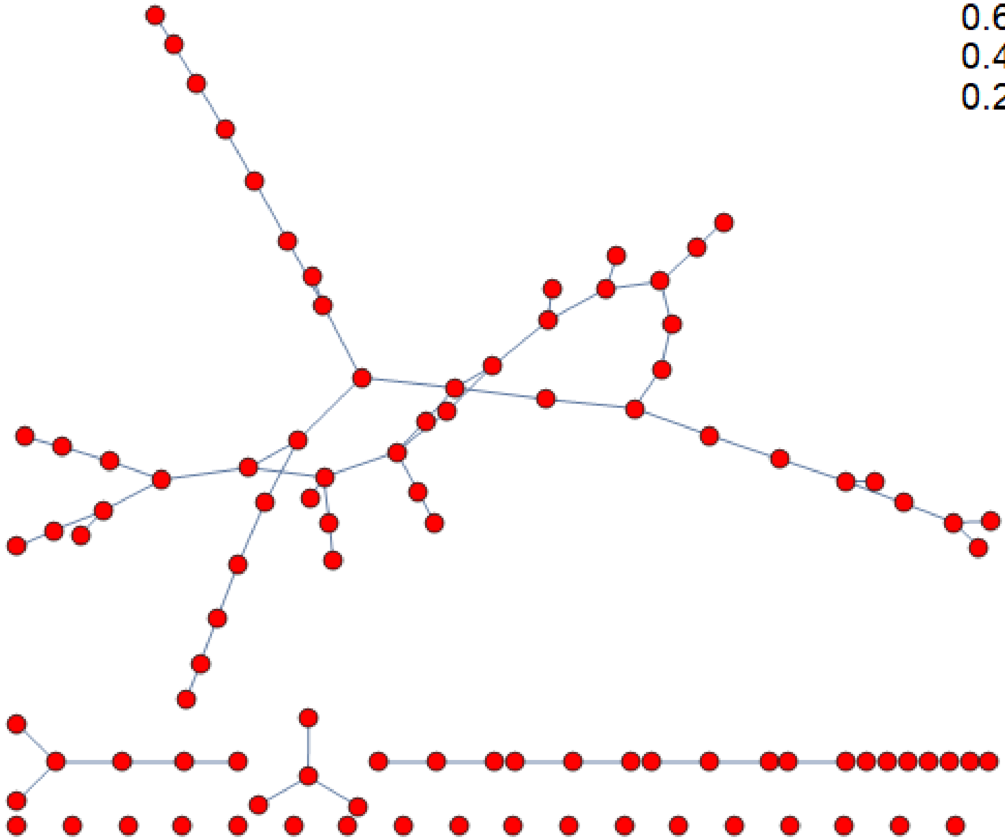




Examples – $\mathrm{ER(N, p)},\ c = 1.5$

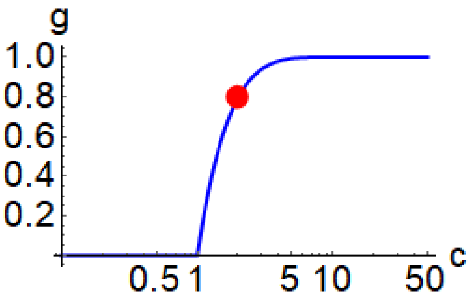
$\mathrm{N} = 100, c = 1.5$ – slightly above the phase transition, with a giant cluster of order N, with many isolated components

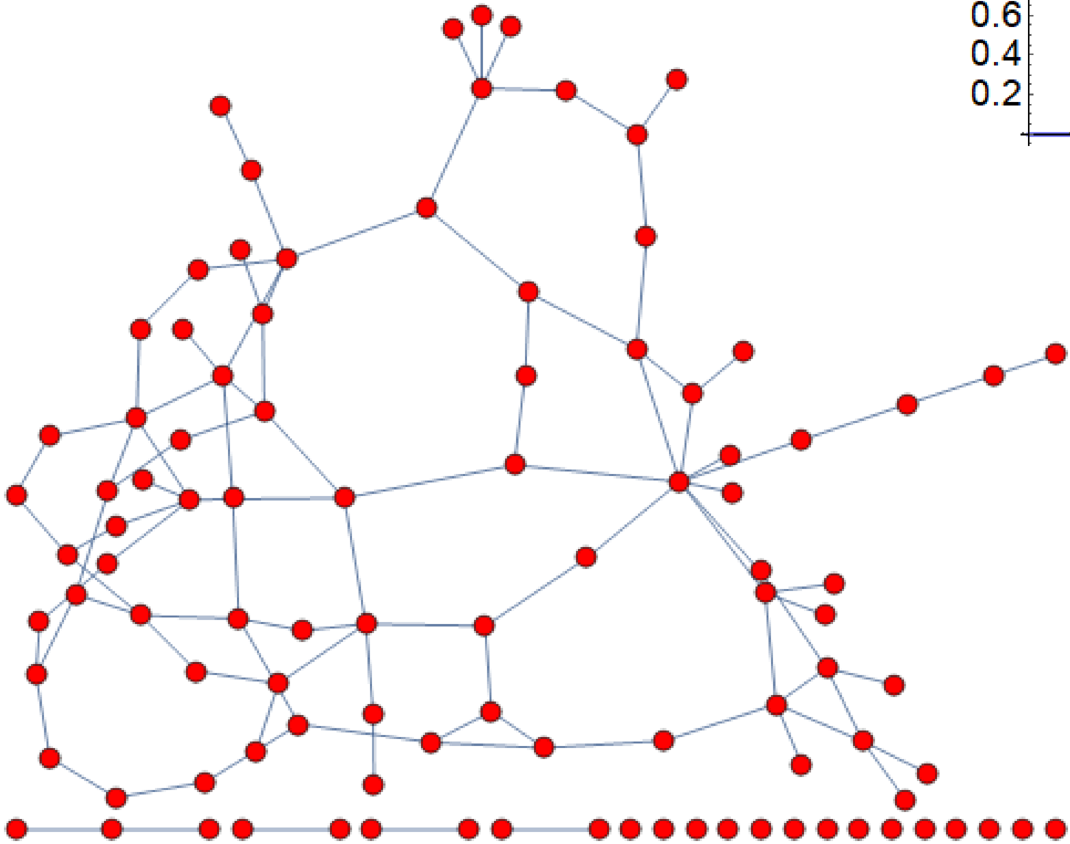




Examples – $\mathrm{ER}(\mathrm{N},\mathrm{p}),\ c = 2$

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

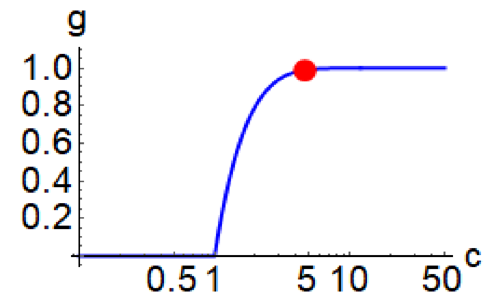


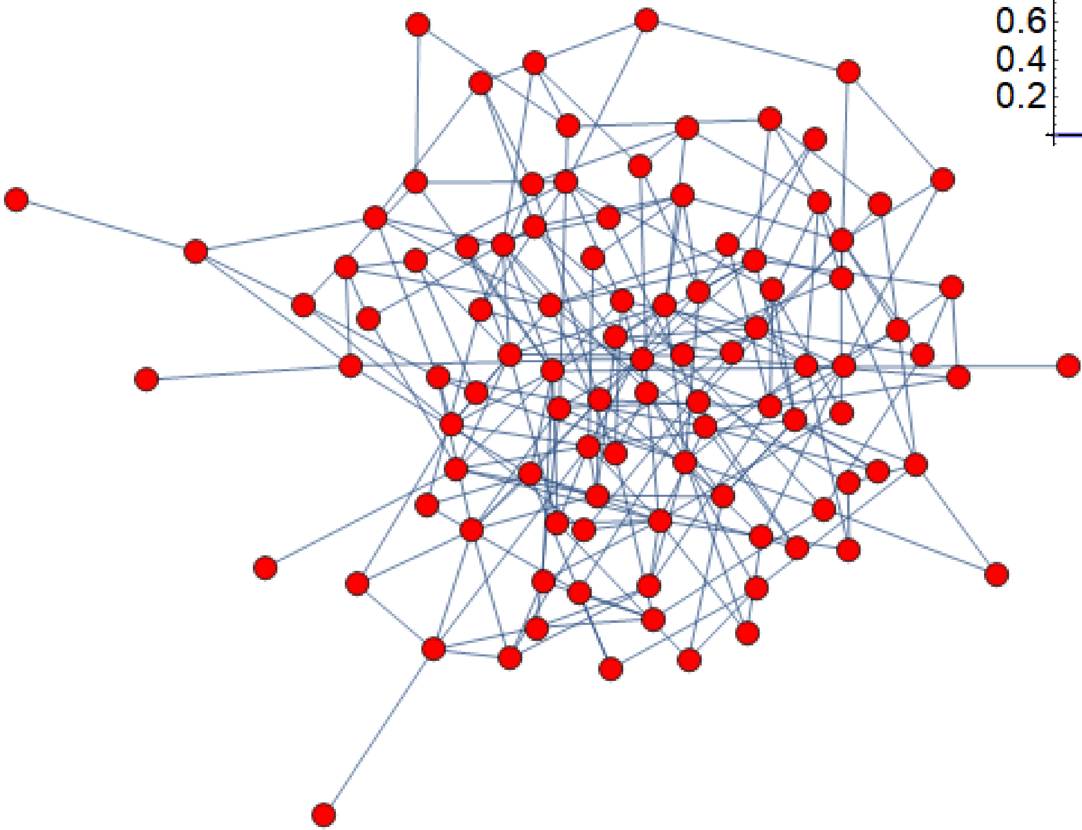


Examples – $\mathrm{ER}(\mathrm{N}, \mathrm{p}),\ c = 4.6$

$\mathrm{N} = 100, c = \ln N$ – one giant cluster with **<u>no</u>** isolated components.

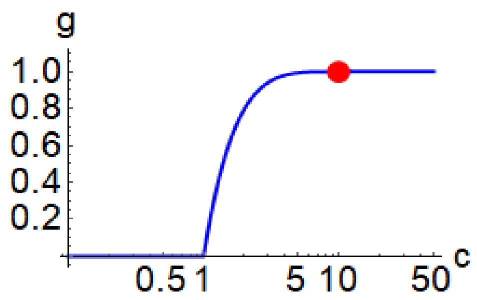
The probability for an isolated node smaller than $1/N$

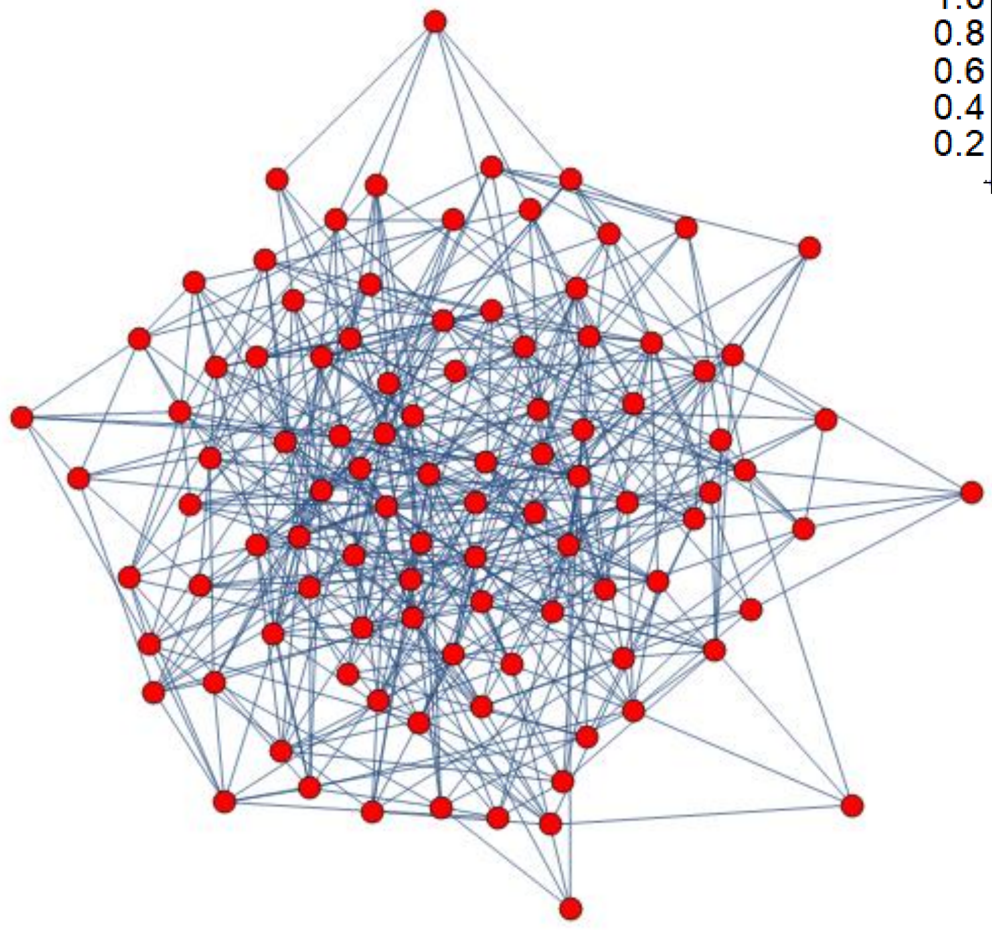




Examples – $\mathrm{ER}(\mathrm{N}, \mathrm{p}),\ c = 10$

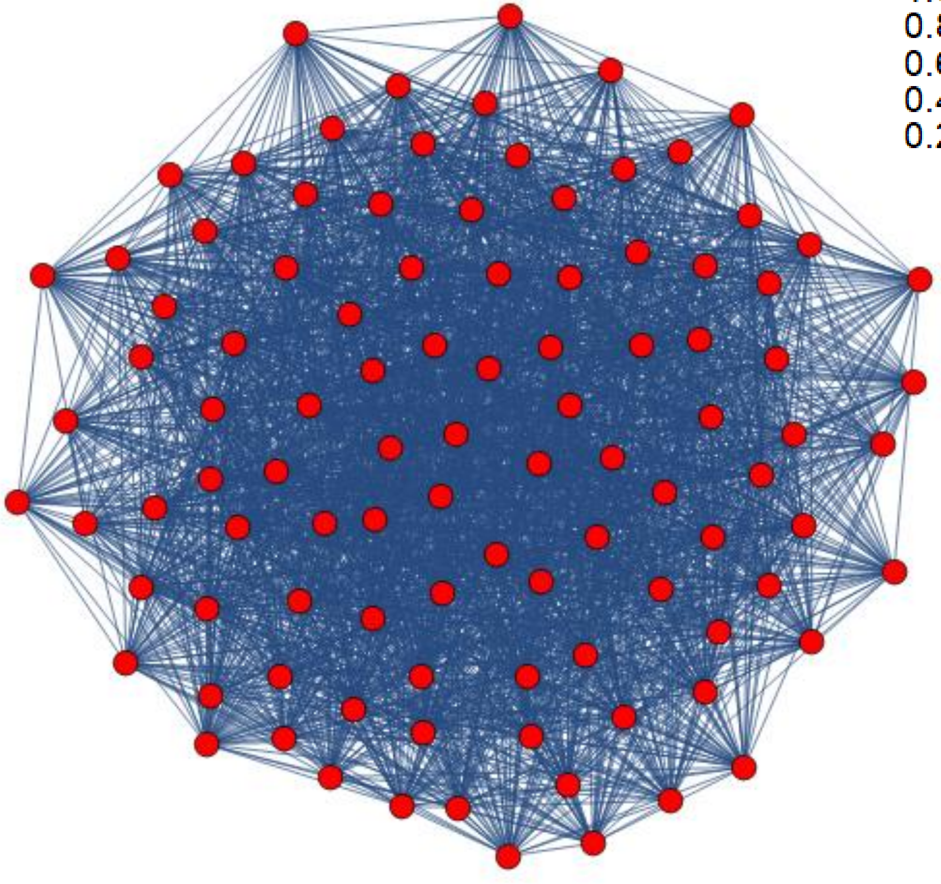
$\mathrm{N} = 100, c = N^{1/2}$ – dense. One giant cluster with no isolated components.

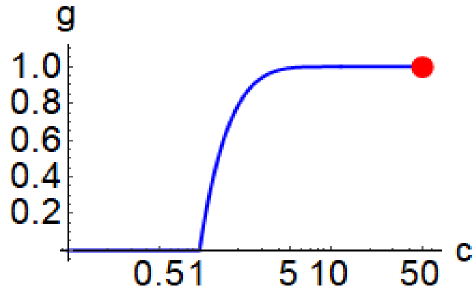




Examples – $\mathrm{ER(N,p)},\ c = 50$

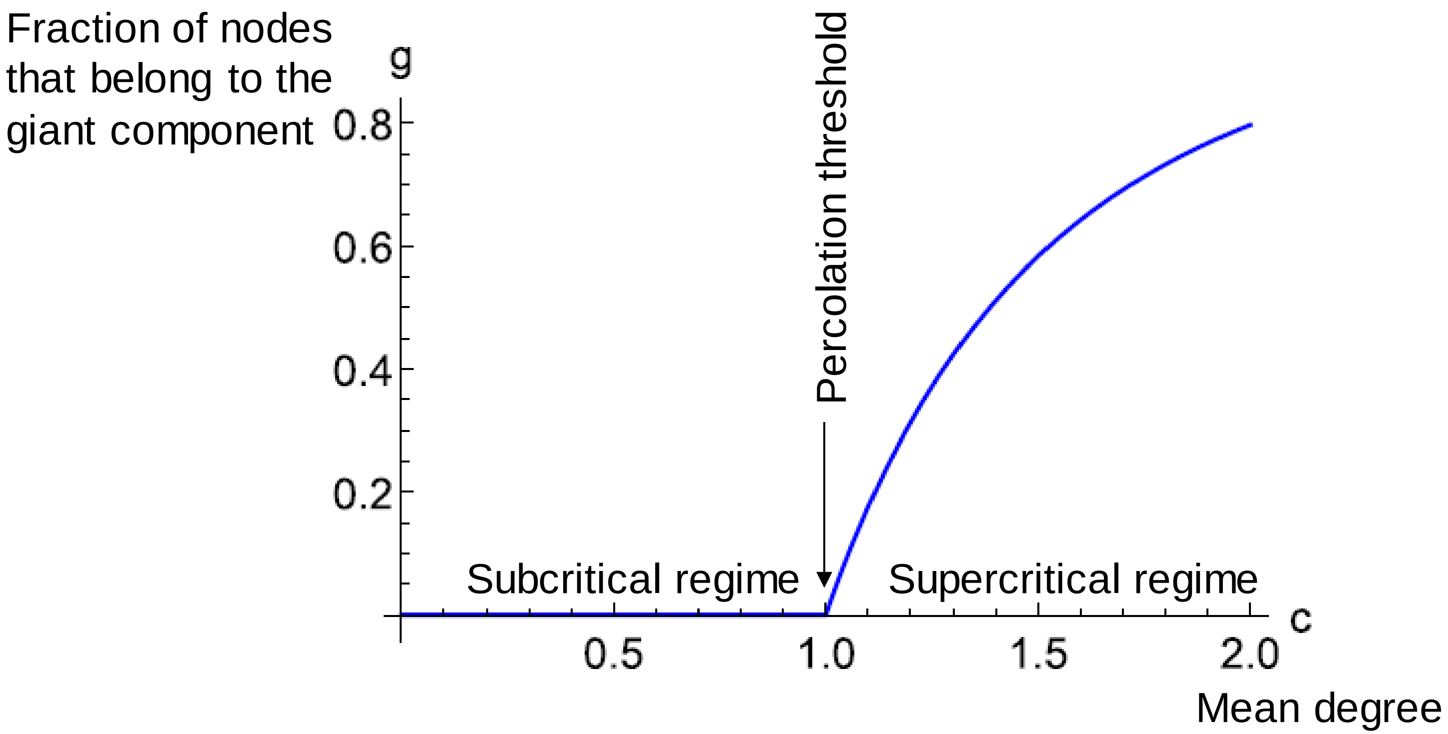
$\mathrm{N} = 100, c = \frac{1}{2}N$ – very dense.





Percolation Transition

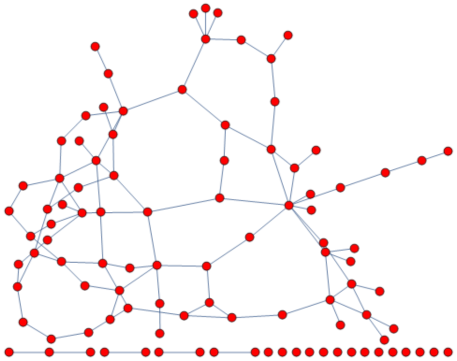
Equilibrium networks undergo a percolation phase transition, below which the network consists of finite components and above it a giant component emerges. For ER networks this happens at $c = 1$. In other ensembles within the configuration it is fixed by the Molloy-Reed criterion: $\frac{\text{excess}}{\text{degree}} = \frac{\langle K^2 \rangle - \langle K \rangle}{\langle K \rangle} = 1$



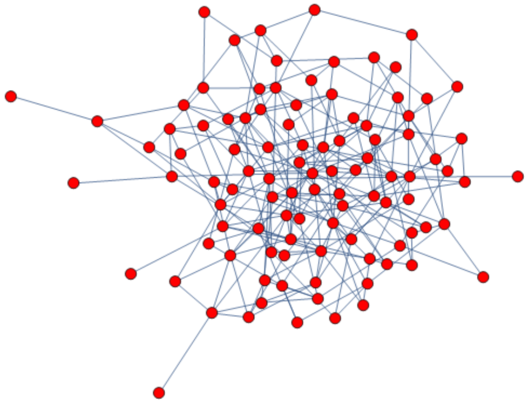
Focus

In this talk we will focus in ER networks in the super-percolating regime, i.e. when the average connectivity is $c > 1$, and even larger than 2.

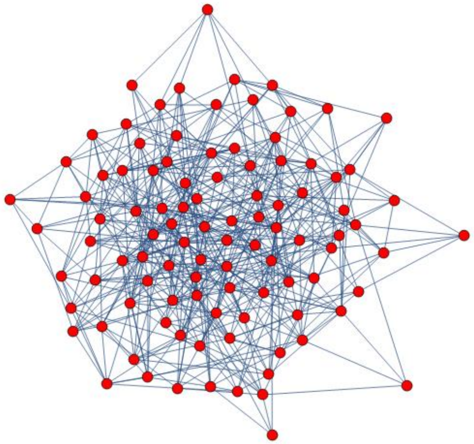
We will however say a few words on subpercolating networks too



$\mathrm{N} = 100, c = 2$



$\mathrm{N} = 100, c = \ln N$



$\mathrm{N} = 100, c = \sqrt{N}$

Metric Properties: Shortest Paths

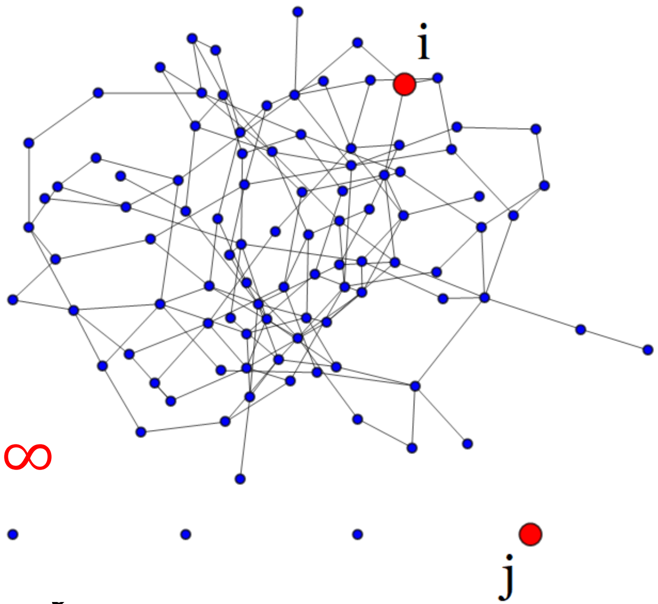
Choose two nodes at random i and j, and consider the different paths connecting them.

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

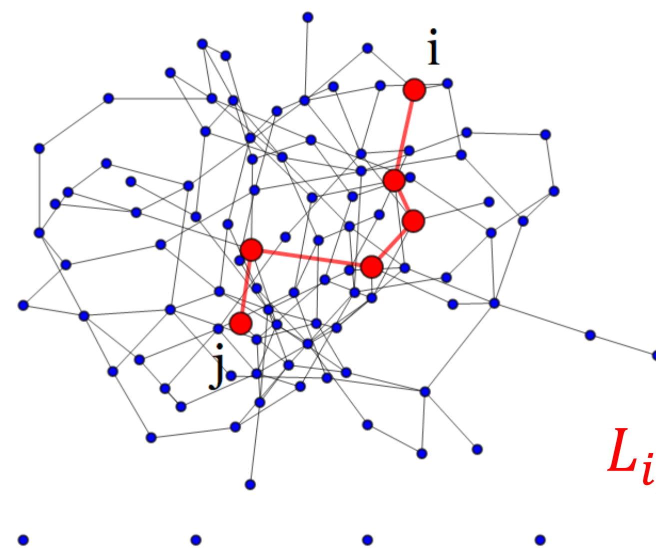
Otherwise, we denote the shortest path length by $L_{ij} < \infty$.

(Note that the shortest path on a finite graph cannot exceed $N - 1$)

In this work we are interested in the distribution of these distances $P(L_{ij})$



$L_{ij} = \infty$



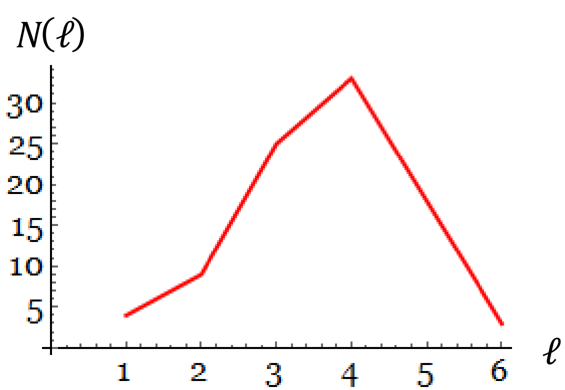
$L_{ij} = 5$

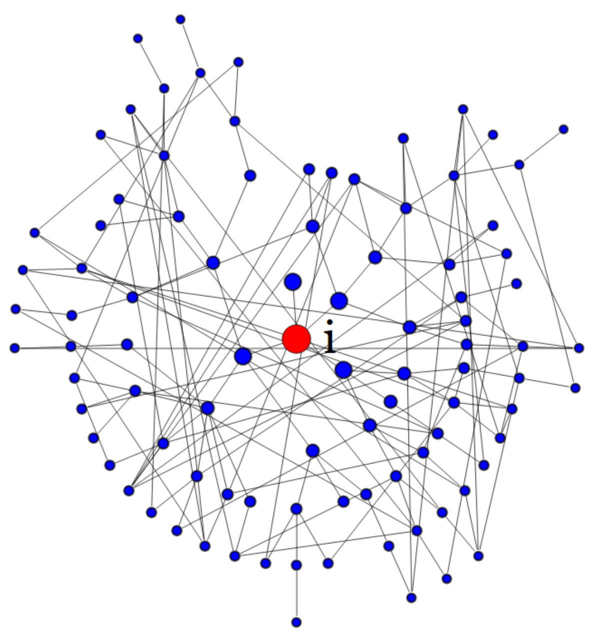
$P(L)$ and the Shell Structure

In this work we are interested in the distribution of these distances $P(L_{ij})$

Note that $P(L_{ij}=\ell)$ is not only the probability that the distance between nodes i and j is equal to ℓ, but also provides the shell structure around the initial node i, namely $\underline{N_i(\ell)\equiv(N-1)\cdot P\left(L_{ij}=\ell\right)}$ counts the number of nodes in the shell located at distance ℓ away from i

$$N_i(\ell)=\{4,9,25,33,18,3;(7)\}$$





$L_{ij}=\infty$

$N_i(\infty)=7$

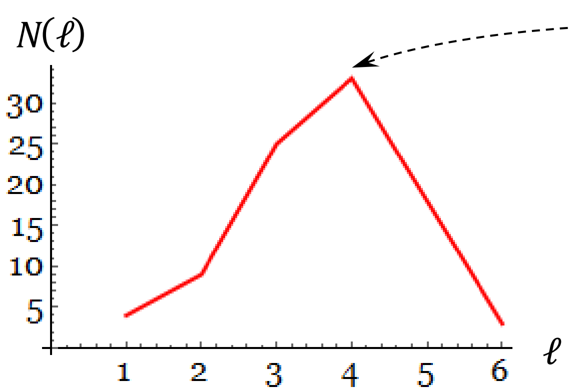
Essentially $N(\ell)$ and $P(L=\ell)$ are the same

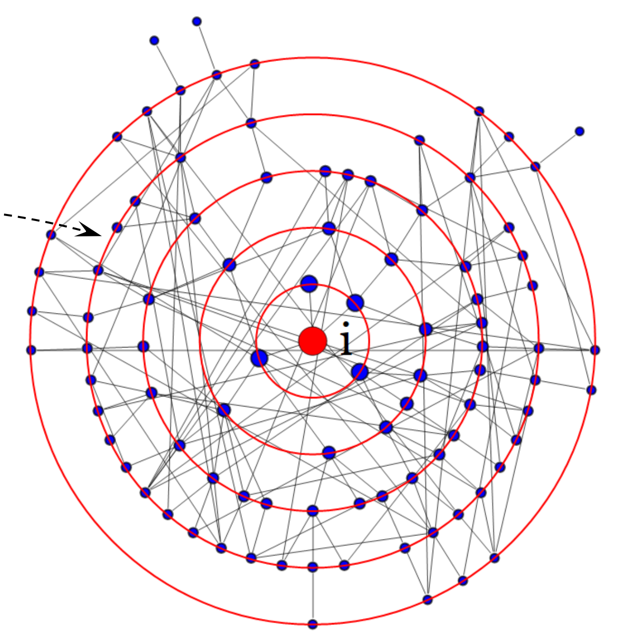
$P(L)$ and the Shell Structure

In this work we are interested in the distribution of these distances $P(L_{ij})$

Note that $P(L_{ij} = \ell)$ is not only the probability that the distance between nodes i and j is equal to ℓ, but also provides the shell structure around the initial node i, namely $\underline{N_i(\ell) \equiv (N-1) \cdot P(L_{ij} = \ell)}$ counts the number of nodes in the shell located at distance ℓ away from i

$N_i(\ell) = \{4, 9, 25, 33, 18, 3; (7)\}$



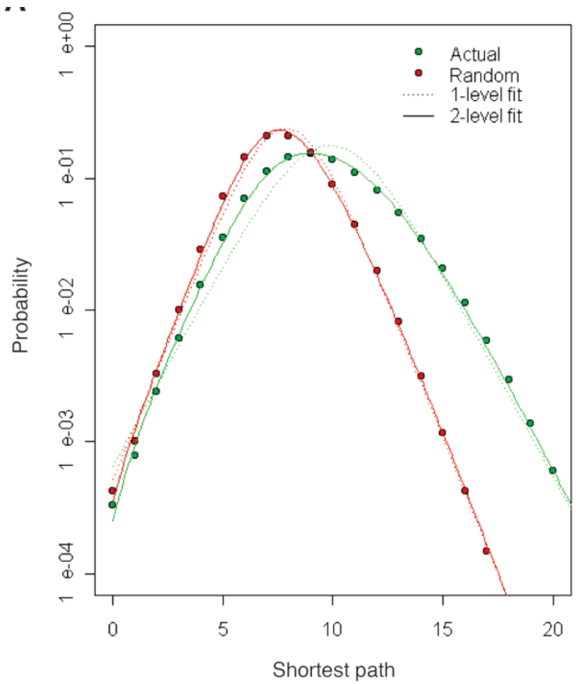


$L_{ij} = \infty$
$N_i(\infty) = 7$

Essentially $N(\ell)$ and $P(L = \ell)$ are the same

Why are they interesting?

<u>Statistical inference</u> - **<u>null models for real networks</u>**. 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.

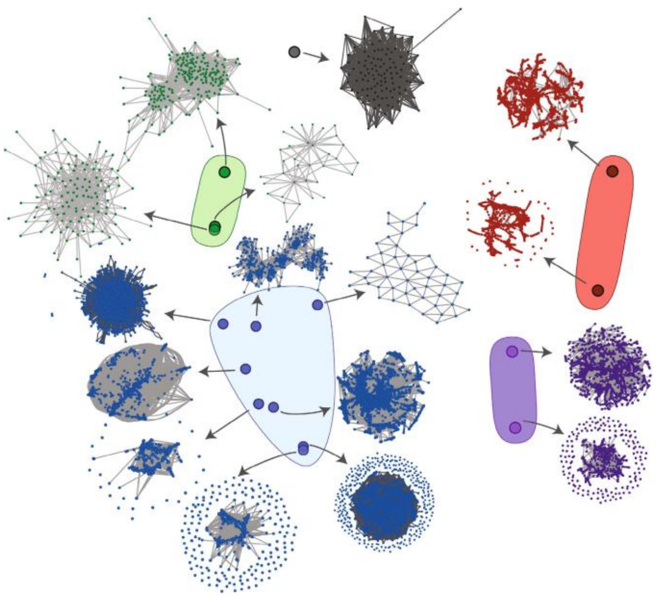


Distances in the empirical networks are longer than the ones in a randomized version of the same network

Giot et al., Science 2003 - A Protein Interaction Map of Drosophila melanogaster

Why are they interesting?

<u>Quantifying distances between graphs</u> – How different are two given networks? Or on the other extreme – are two graphs isomorphic? Since the DSPL is a strong fingerprint of a network, and also computationally simply to obtain, it can serve as a useful characteristic to differentiate between graphs.



Multidimensional scaling map of the set of real networks performed over the averaged D-values

Schreiber et al., Nat. Comm. 2016 - Quantification of network structural dissimilarities

Shortest Paths – known results

Known results are mainly for the equilibrium sparse networks $p = c/N$

The **mean distance** in this case is

$$\langle L \rangle = \sum_{\ell} \ell \cdot P_N(L = \ell) \simeq \frac{\ln N}{\ln c} + O(1)$$

which exhibits/explains the so-called "small-world phenomenon".

The **diameter** in these cases typically scales ~ in the same way.

In scale-free networks this is typically "ultra-small world",

i.e., $\langle L \rangle \sim \ln \ln N$ [Cohen+Havlin, 2003]

The full Distribution of Shortest Path Length (**DSPL**) is nevertheless of great interest, yet not received enough attention. When studied, it gained limited success even for the simple ER case.

Furthermore, the potential utility of this **DSPL** has not been much explored.

Shortest Paths – known results

The full Distribution of Shortest Path Length (**DSPL**) is nevertheless of great interest, yet not received enough attention. When studied, it gained limited success even for the simple ER case.

- Newman, Strogatz & Watts, *Phys. Rev. E* **64**, 026118 (2001).
- Dorogovtsev, Mendes & Samukhin, *Nucl. Phys. B* **653**, 307 (2003).
- van der Hofstad, Hooghiemstra & van Mieghem, *Rand. Struct. Algo.* **27**, 76 (2005).
- Blondel, Guillaume, Hendrickx & Jungers, *Phys. Rev. E* **76**, 066101 (2007).
- van der Hofstad, Hooghiemstra & Znamenski, *Electron. J. Probab.* **12**, 703 (2007).
- van der Hofstad & Hooghiemstra, *Math. Phys.* **49**, 125209 (2008).
- van der Esker, van der Hofstad & Hooghiemstra, *J. Stat. Phys.* **133**, 169 (2008).
- Shao, Buldyrev, Cohen, Kitsak, Havlin & Stanley, *Europhys. Lett.* **84**, 48004 (2008).
- Shao, Buldyrev, Braunstein, Havlin & Stanley, *Phys. Rev. E* **80**, 036105 (2009).
- Goldental, Vardi, Sardi, Sabo & Kanter, *Front. Neural Circuits* **9**, 65 (2015).
- Melnik & Gleeson, arXiv:1604.05521 (2016).
- Jackson & Patil, *J. Stat. Phys.* **189**, 40 (2022).

Shortest Paths – our results

- E. Katzav, M. Nitzan, D. ben-Avraham, P.L. Krapivsky, R. Kühn, N. Ross and O. Biham, Analytical results for the DSPL in random networks, *EPL* **111**, 26006 (2015).
- M. Nitzan, E. Katzav, R. Kühn and O. Biham, Distance distribution in configuration model networks, *Phys. Rev. E* **93**, 062309 (2016).
- C. Steinbock, O. Biham and E. Katzav, The distribution of shortest path lengths in a class of node duplication network models, *Phys. Rev. E* **96**, 032301 (2017).
- H. Bonneau, A. Hassid, O. Biham, R. Kühn and E. Katzav, Distribution of shortest cycle lengths in random networks, *Phys. Rev. E* **96**, 062307 (2017).
- I. Tishby, O. Biham, E. Katzav and R. Kühn, Revealing the Micro-Structure of the Giant Clusters in Random Graph Ensembles, *Phys. Rev. E* **97**, 042318 (2018).
- E. Katzav, O. Biham and A. Hartmann, The distribution of shortest path lengths in subcritical Erdős-Rényi networks, *Phys. Rev. E* **98**, 012301 (2018).
- C. Steinbock, O. Biham and E. Katzav, Exact results for directed random networks that grow by node duplication, *Eur. Phys. J. B* **92**, 130 (2019).
- I. Tishby, O. Biham, R. Kühn and E. Katzav, The mean and variance of the distribution of shortest path lengths of random regular graphs, *J. Phys. A* **55**, 265005 (2022).
- B. Budnick, O. Biham and E. Katzav, The distribution of shortest path lengths on trees of a given size in subcritical Erdős-Rényi networks, submitted to Phys. Rev. E (2023)

Shortest Paths – our results

- E. Katzav, M. Nitzan, D. ben-Avraham, P.L. Krapivsky, R. Kühn, N. Ross and O. Biham, Analytical results for the DSPL in random networks, *EPL* **111**, 26006 (2015).

EPL (Europhysics Letters)
Authors Access

EPL Portal | Referees Access | Editors Access

Manuscript status

Reference:	**G35008**
Author(s):	Eytan Katzav, Mor Nitzan, Daniel ben-Avraham, P. L. Krapivsky, Reimer Kühn, Nathan Ross and Ofer Biham
Submission date:	21/04/15
Date accepted:	20/07/15
PACS number(s):	64.60.aq 89.75.Da
Co-editor:	Professor Jose A. Cuesta
Title:	Analytical results for the distribution of shortest path lengths in random networks
Status:	**Manuscript accepted**

History of the paper

Date:	**Event:**
22 Apr 2015	Manuscript received by Editorial Office
22 Apr 2015	Manuscript sent to Co-editor

Shortest Paths – our results

22 Apr 2015	Manuscript sent to Referee1
22 Apr 2015	Manuscript sent to Referee2
29 Apr 2015	Reminded Referee1
29 Apr 2015	Reminded Referee2
29 Apr 2015	Manuscript sent to Referee3
29 Apr 2015	Referee2 could not report
30 Apr 2015	Manuscript sent to Referee4
30 Apr 2015	Referee3 could not report
30 Apr 2015	Referee4 could not report
30 Apr 2015	Manuscript sent to Referee5
04 May 2015	Referee5 could not report
04 May 2015	Manuscript sent to Referee6
04 May 2015	Referee6 could not report
04 May 2015	Manuscript sent to Referee7
06 May 2015	2nd reminder to Referee1
11 May 2015	Reminded Referee7
12 May 2015	Referee7 sent a report
12 May 2015	Manuscript sent to Referee8
12 May 2015	Referee8 could not report
12 May 2015	Manuscript sent to Referee9
13 May 2015	Manuscript sent to Referee10
13 May 2015	Referee9 could not report
20 May 2015	Reminded Referee10
27 May 2015	2nd reminder to Referee10
29 May 2015	Manuscript sent to Referee11
29 May 2015	Manuscript sent to Referee12
05 Jun 2015	Reminded Referee11
05 Jun 2015	Reminded Referee12
12 Jun 2015	2nd reminder to Referee11
12 Jun 2015	2nd reminder to Referee12
19 Jun 2015	Manuscript sent to Referee13
19 Jun 2015	Manuscript sent to Referee14
24 Jun 2015	Referee13 could not report
24 Jun 2015	Manuscript sent to Referee15
26 Jun 2015	Referee14 could not report
26 Jun 2015	Manuscript sent to Referee16
29 Jun 2015	Referee16 could not report
29 Jun 2015	Manuscript sent to Referee17
30 Jun 2015	Referee17 could not report
30 Jun 2015	Manuscript sent to Referee18
30 Jun 2015	Manuscript sent to Referee19
30 Jun 2015	Referee18 could not report

Shortest Paths – our results

01 Jul 2015	Reminded Referee15
07 Jul 2015	Reminded Referee19
08 Jul 2015	2nd reminder to Referee15
13 Jul 2015	Manuscript sent to Referee20
15 Jul 2015	2nd reminder to Referee19
16 Jul 2015	Manuscript sent to Referee21
17 Jul 2015	Referee21 could not report
20 Jul 2015	Referee20 could not report
22 Jul 2015	Manuscript Accepted

The editor, Prof. Jose A. Cuesta, simply accepted the paper after 21 referees declined to report …

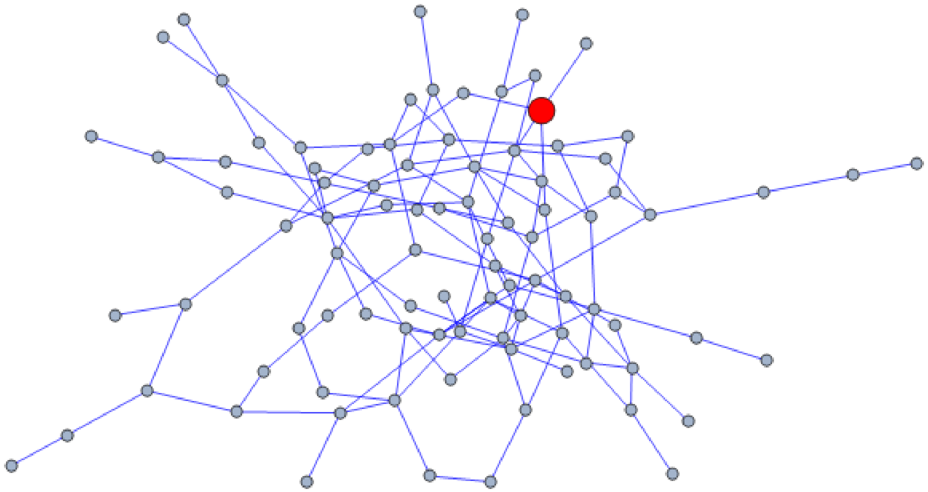
Importance: epidemic spreading

We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

$t = 0$



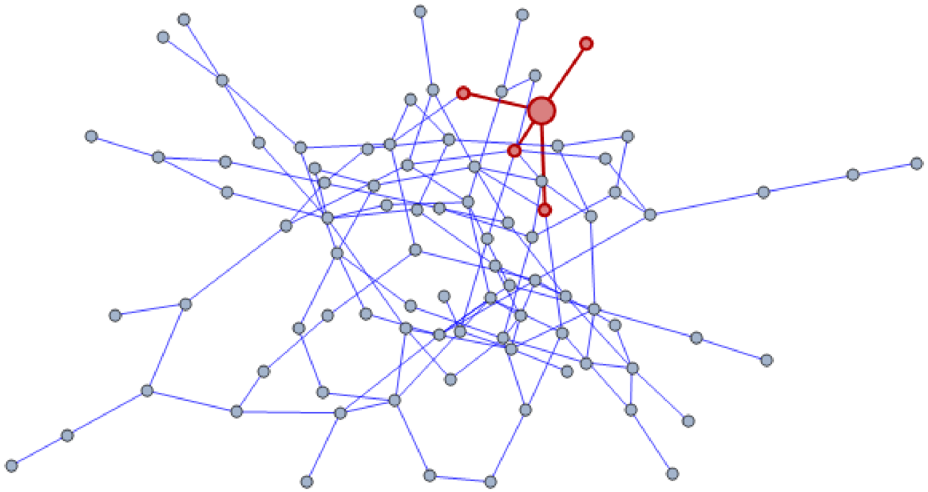
Importance: epidemic spreading

We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

$t = 1$



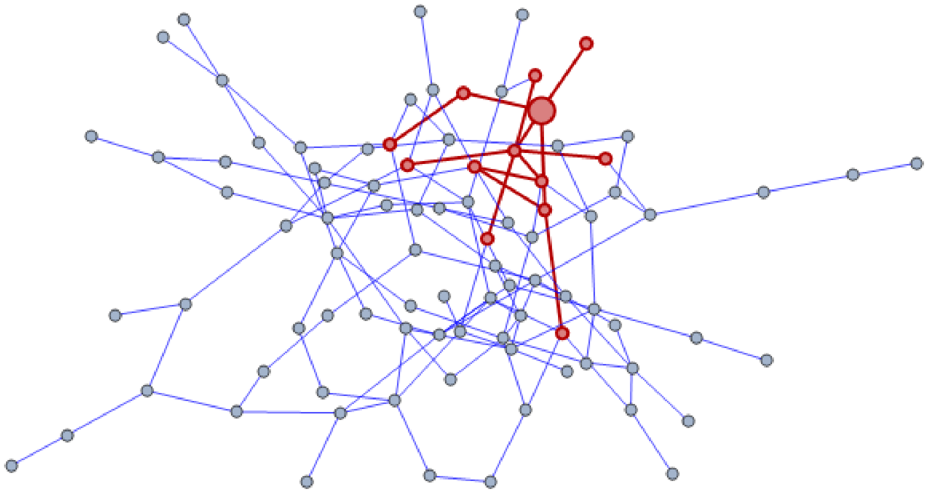
Importance: epidemic spreading

We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

$t = 2$



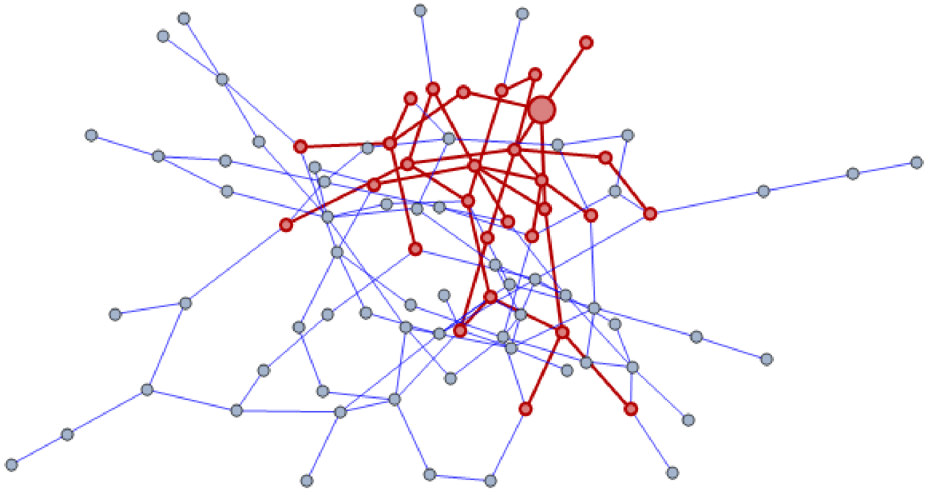
Importance: epidemic spreading

We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

$t = 3$



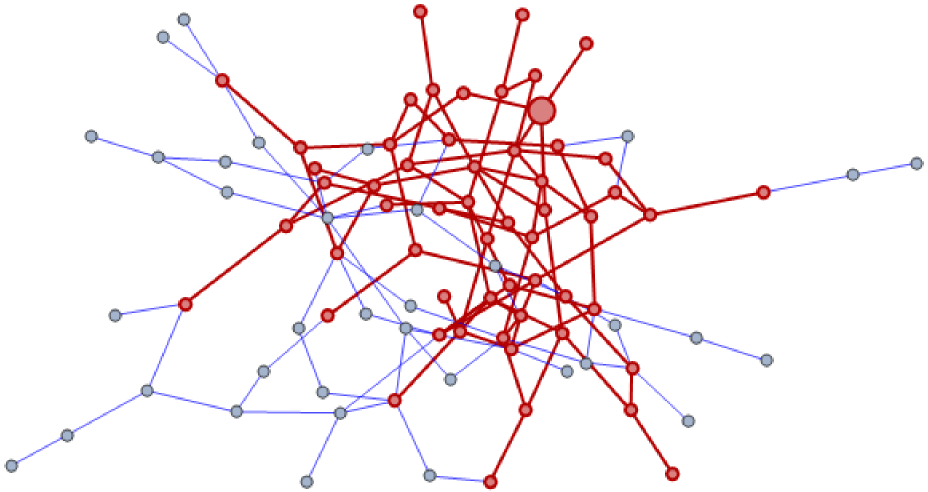
Importance: epidemic spreading

We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

$t = 4$



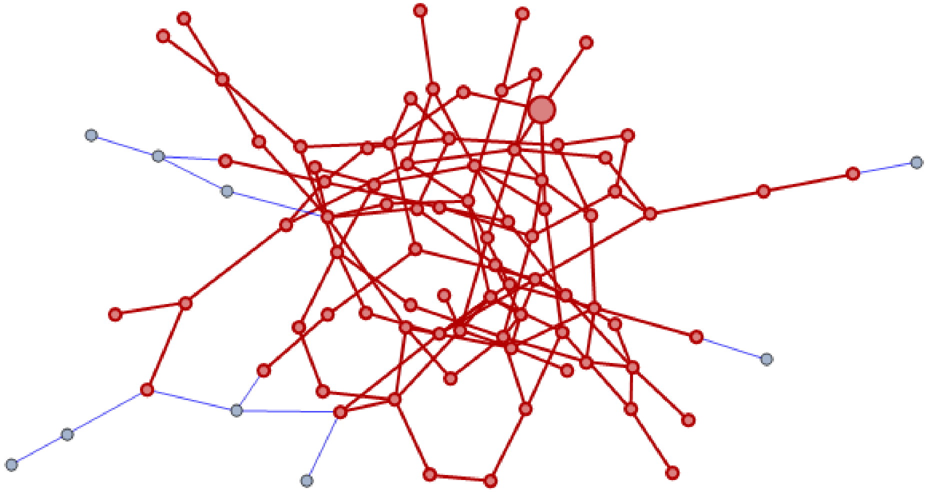
Importance: epidemic spreading

We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

$t = 5$



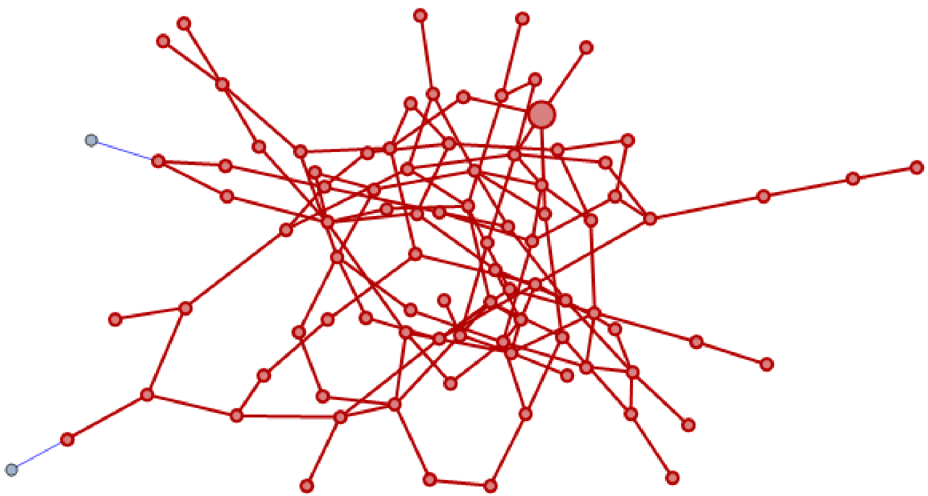
Importance: epidemic spreading

We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

$t = 6$



Importance: epidemic spreading

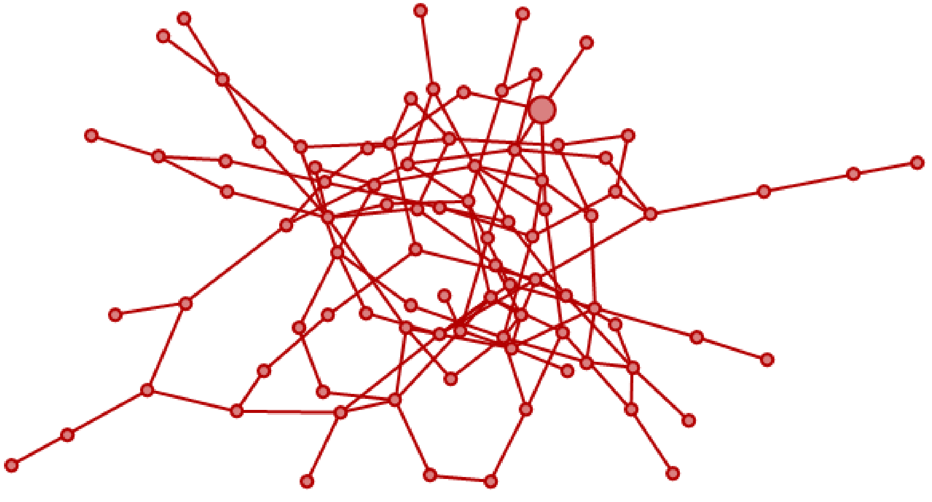
We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

Actually, the DSPL provides a detailed temporal account of this process…

$t = 7$



Importance: epidemic spreading

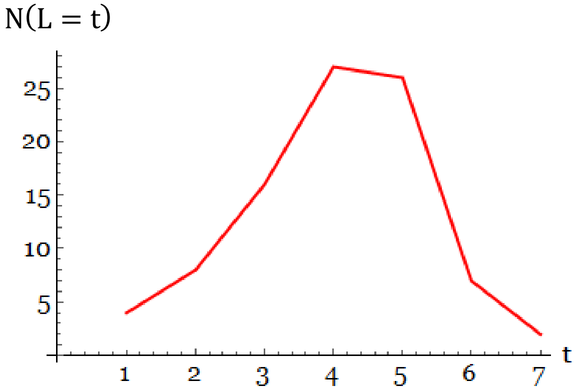
We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

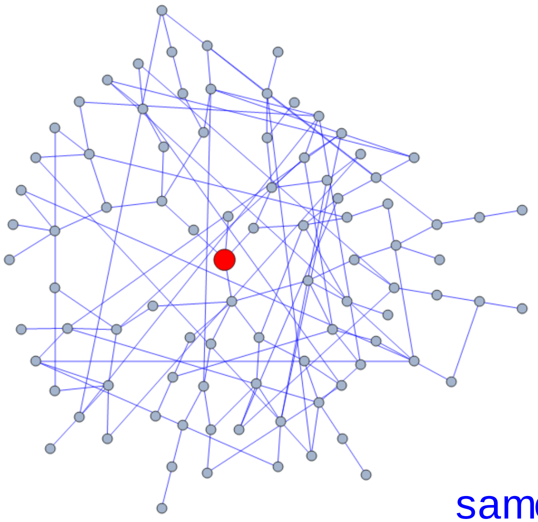
This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

Actually, the DSPL provides a detailed temporal account of this process…

$t = 0$





same networks

Importance: epidemic spreading

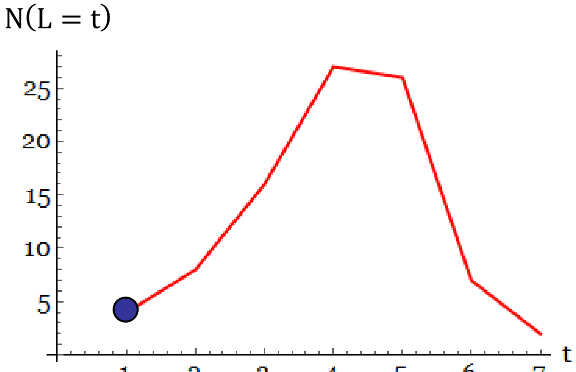
We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

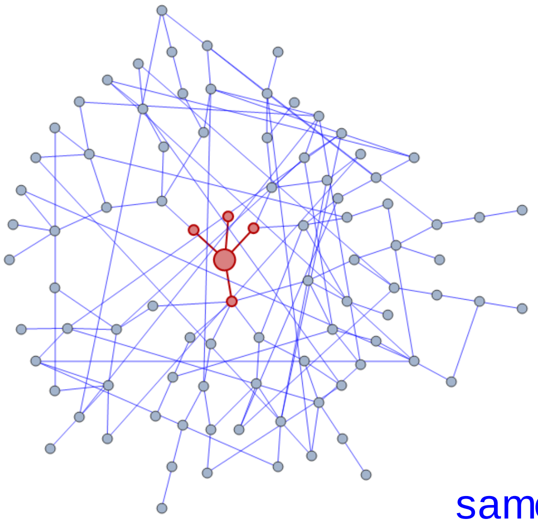
This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

Actually, the DSPL provides a detailed temporal account of this process…

$t = 1$





same networks

Importance: epidemic spreading

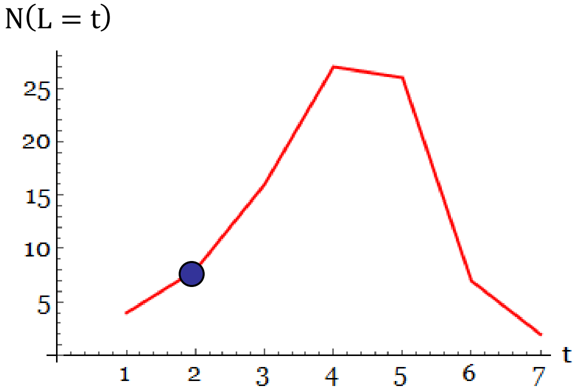
We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

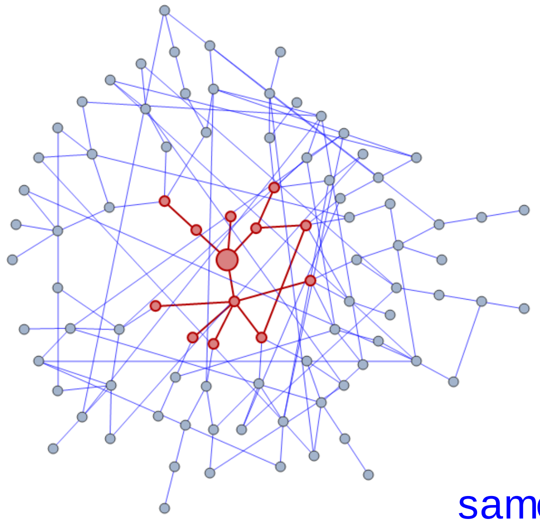
This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely ($p_i = 1$):

Actually, the DSPL provides a detailed temporal account of this process…

$t = 2$





same networks

Importance: epidemic spreading

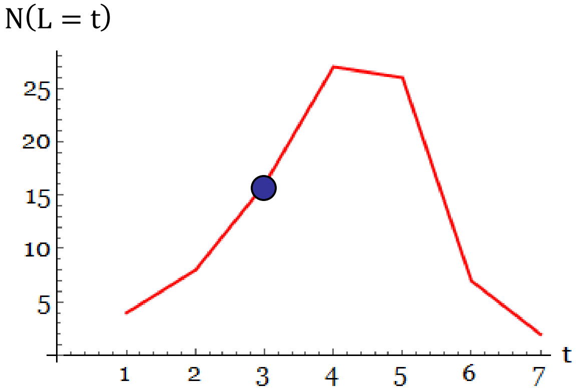
We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

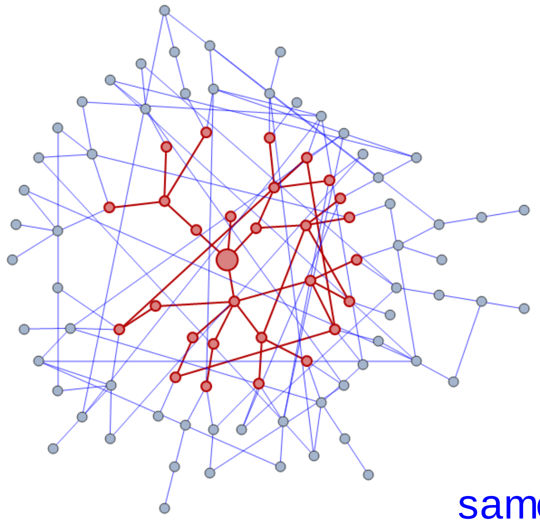
This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

Actually, the DSPL provides a detailed temporal account of this process…

$t = 3$





same networks

Importance: epidemic spreading

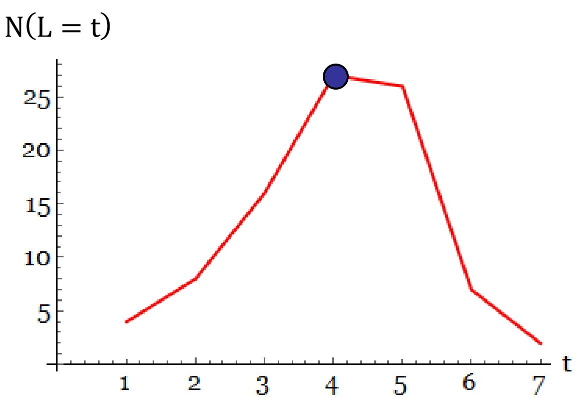
We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

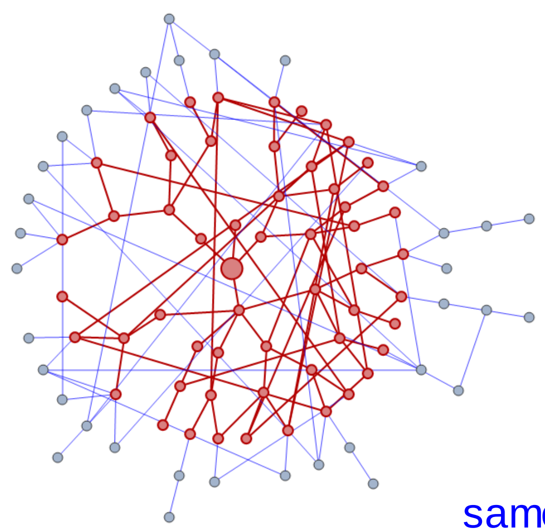
This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

Actually, the DSPL provides a detailed temporal account of this process…

$$t = 4$$





same networks

Importance: epidemic spreading

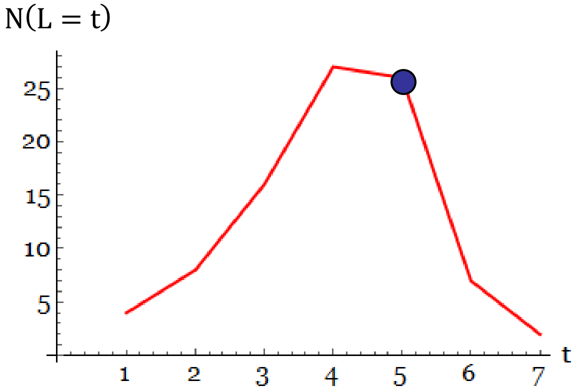
We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

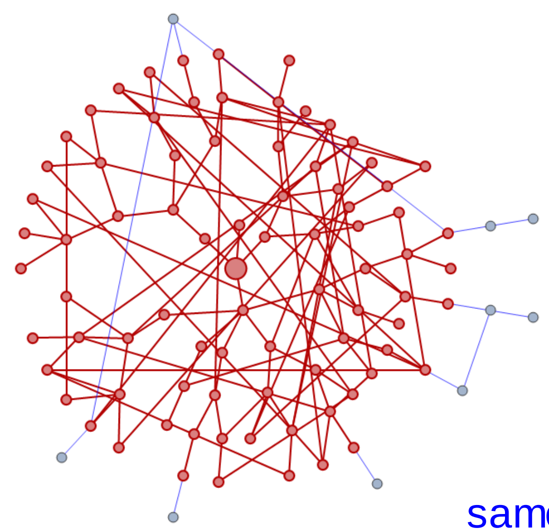
This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

Actually, the DSPL provides a detailed temporal account of this process…

$t = 5$





same networks

Importance: epidemic spreading

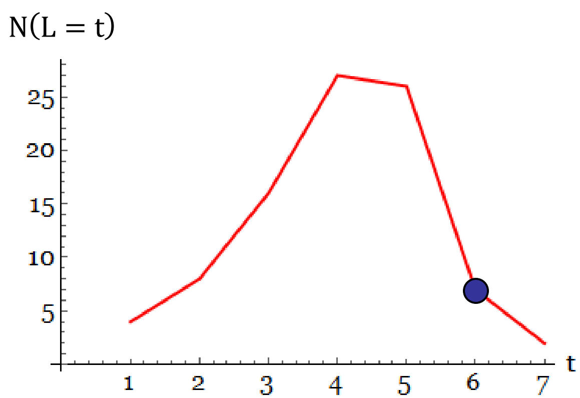
We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

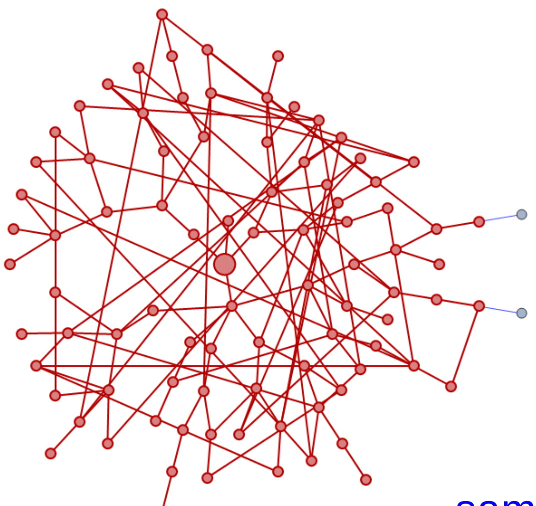
This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

Actually, the DSPL provides a detailed temporal account of this process…

$t = 6$





same networks

Importance: epidemic spreading

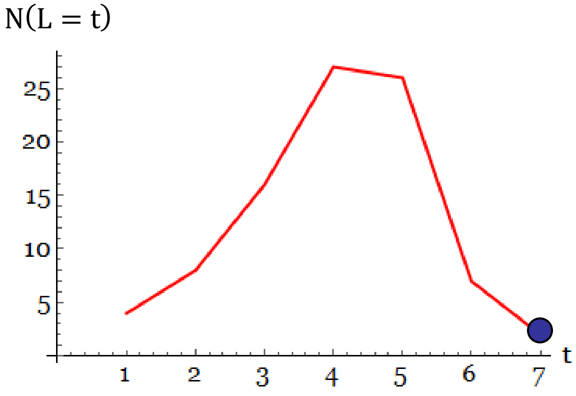
We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

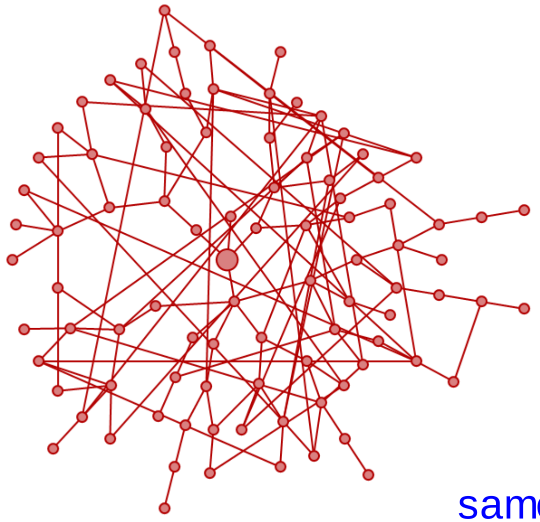
This is irrespective of how $P(L)$ is actually calculated or measured!

As an illustration, consider a very contagious epidemic spreading on a network, such that an infected node infects all its neighbours surely $(p_i = 1)$:

Actually, the DSPL provides a detailed temporal account of this process…

$t = 7$





same networks

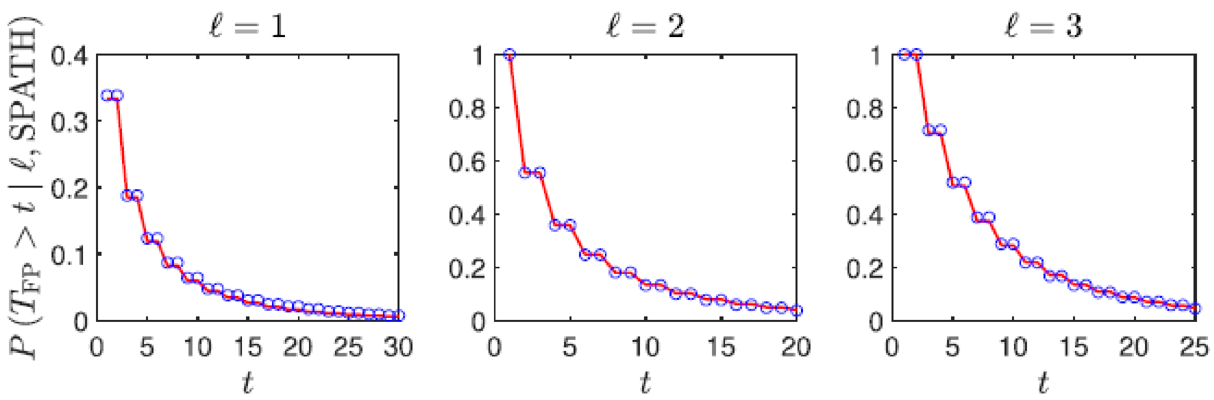
Importance: First Passage Time

We claim that knowing $P(L)$ is actually a very powerful piece of information, for example, in dynamical processes on networks.

This is irrespective of how $P(L)$ is actually calculated or measured!

Another important example is the First Passage Time problem, namely the first time that a RW starting at some node i reaches a specific node j. The initial distance L_{ij} between is important. It is essential for $P(\text{SPATH})$.

$$P(T_{\text{FP}} = t|\ell, \text{SPATH}) = \begin{cases} \frac{\ell}{t}\binom{t}{\frac{t+\ell}{2}}\left(1-\frac{1}{c}\right)^{\frac{t+\ell}{2}}\left(\frac{1}{c}\right)^{\frac{t-\ell}{2}} & t-\ell \text{ even} \\ 0 & t-\ell \text{ odd.} \end{cases}$$



Tishby, Biham & Katzav, Stat. Mech. 113403 (2022).

Importance: radial expansion…

If the epidemic is less contagious, with $p_i < 1$, we get a **diffusive process** on the network $\partial_t n = \Delta n$, yet the DSPL still provides its (radial) backbone.

Actually, it also provides the radial propagation of waves, obeying $\partial_{tt}\psi = \Delta\psi$ - namely the DSPL describes the **light-cone** of emitted signals.

More generally, this is the quantity which is needed to perform a separation of variables of the Laplacian and gives the radial part.

Think of $d^3r = r^2 dr\,\mathrm{d}\Omega(\theta,\phi)$, then $P(\ell)$ or $N(\ell)$ is analogous to r^2 (or r^{D-1}).

In a sense, this determines Coulomb's law on a network.

From a different perspective, random networks are in a sense infinite dimensional since typical distances in D-dimensions scale as $\langle L\rangle \sim N^{1/D}$, while in random networks $\langle L\rangle \sim \ln N$, which suggests $D \to \infty$.

Other metric properties

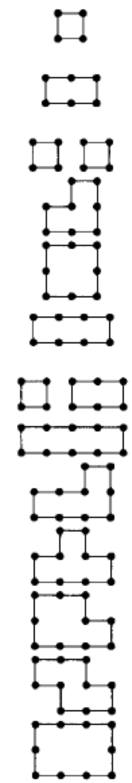
The DSPL is only one example of a family of distributions of metric properties on networks which are of great importance.

Another example is the **distribution of shortest cycle lengths** $P_{DSCL}(L = \ell)$.

Cycles play an important role in the study of critical phenomena (e.g. the Ising model) on networks using high temperature expansions as well as in dynamical processes such as the first return of diffusive particles and re-infection in epidemic models.

Shortest cycles create the strongest feedback.

Order	Contributing graphs	Count
v^4		N
v^6		$2N$
v^8		$N(N-5)/2$
		$4N$
		N
		$2N$
v^{10}		$2N(N-8)$
		$2N$
		$8N$
		$4N$
		$8N$
		$4N$
		$2N$



Yeomans, 1992

… Shortest Cycles

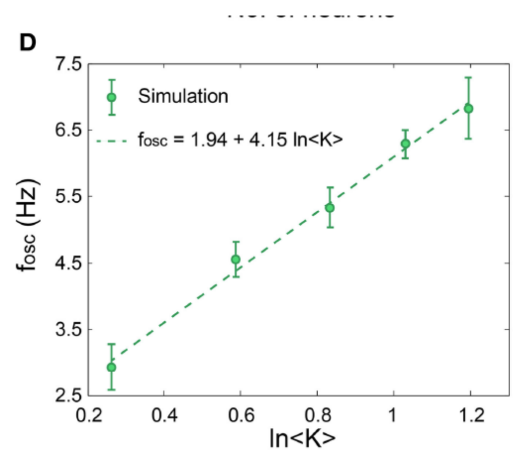
frontiers in Neural Circuits

ORIGINAL RESEARCH
published: 30 October 2015
doi: 10.3389/fncir.2015.00065

Broadband macroscopic cortical oscillations emerge from intrinsic neuronal response failures

Amir Goldental[1†], Roni Vardi[2†], Shira Sardi[1,2], Pinhas Sabo[1] and Ido Kanter[1,2]*

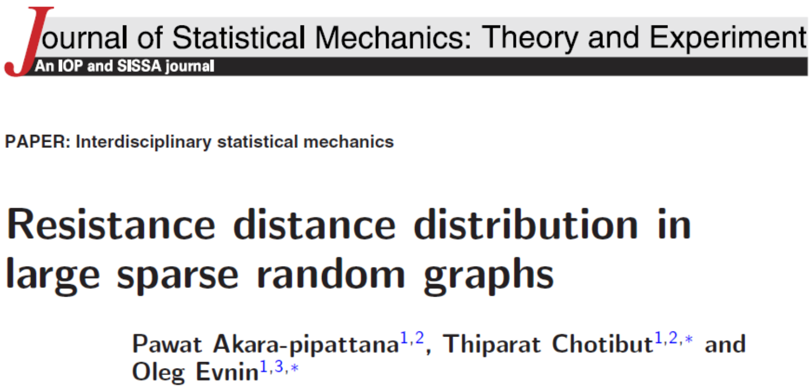
[1] Department of Physics, Bar-Ilan University, Ramat-Gan, Israel, [2] Gonda Interdisciplinary Brain Research Center, Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel



Shortest cycles create the strongest feedback and hence set the time-scale for broadband macroscopic oscillations in the brain.

This is an example for a system where a random network actually demonstrates how robust/universal the phenomenon of large scale oscillations is (compared to experiments and simulations).

Resistance Distance



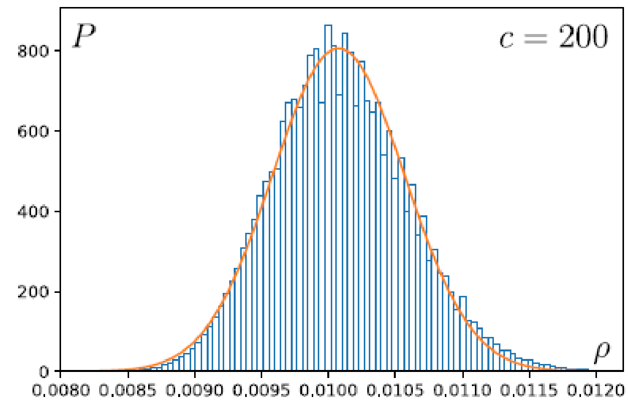
Journal of Statistical Mechanics: Theory and Experiment

An IOP and SISSA journal

PAPER: Interdisciplinary statistical mechanics

Resistance distance distribution in large sparse random graphs

Pawat Akara-pipattana[1,2], Thiparat Chotibut[1,2,*] and Oleg Evnin[1,3,*]



Resistance distance – is simply the electric resistance between two nodes i and j in the network, assuming that all edges are of 1Ω. Resistance distances capture properties of all possible paths between i and j. Interestingly, R_{ij} is a proper metric in the graph, namely obeys all axioms, including the triangle inequality.

This can be expressed through the inverse of the graph Laplacian and is therefore closely related to diffusion and random walks on graphs.

P. Akara-pipattana, T. Chotibut and O. Evnin, Resistance distance distribution in large sparse random graphs, *J. Stat. Mech.* 033404 (2022).

P. Akara-pipattana and O. Evnin, Random matrices with row constraints and eigenvalue distributions of graph Laplacians, *J. Phys. A* **56**, 295001 (2023).

The DSPL of RRG: naïve approach

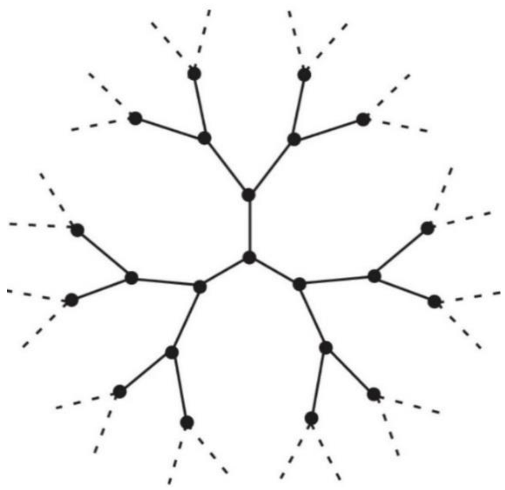
Let's start with a naïve approach to calculate the DSPL of RRG's, essentially based on the Cayley tree:

$$N_\ell = 1, c, c(c-1), c(c-1)^2, \ldots$$

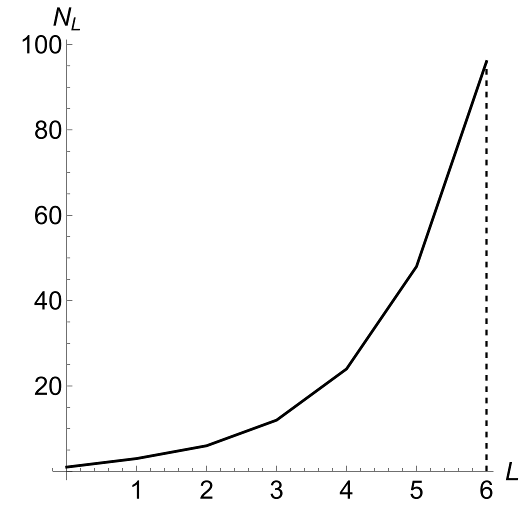
Meaning that there is an exponential growth in the volume (hyperbolic geometry)

Infinite RRGs ($N \to \infty$)

Can't go on forever!



Tree graphs ($\mathrm{c} = 3$)



The DSPL of RRG: naïve approach

Let's start with a naïve approach to calculate the DSPL of RRG's, essentially based on the Cayley tree:

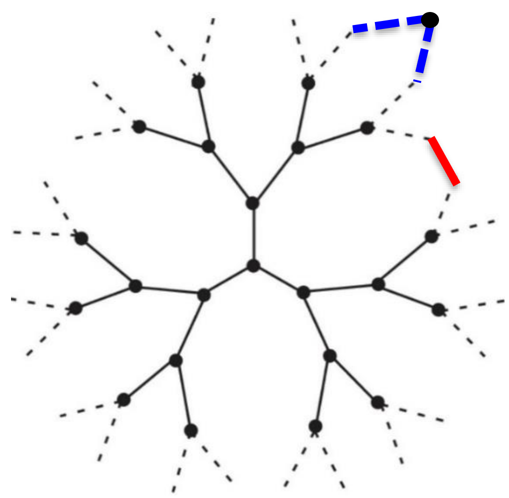
$$N_\ell = 1, c, c(c-1), c(c-1)^2, \ldots$$

Meaning that there is an exponential growth in the volume (hyperbolic geometry)

Infinite RRGs ($N \to \infty$)

Can't go on forever:

two things can happen



Tree graphs $(\mathrm{c} = 3)$

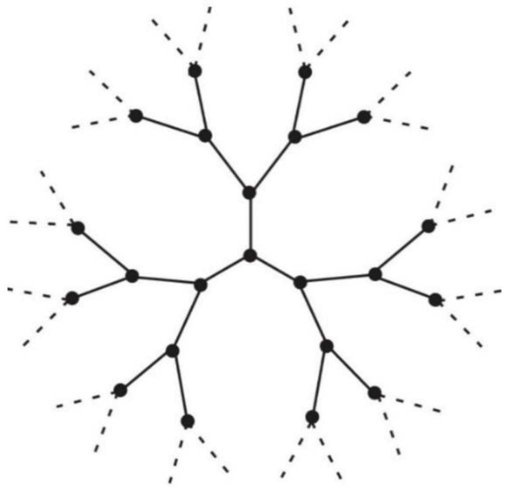
The DSPL of RRG: naïve approach

Let's start with a naïve approach to calculate the DSPL of RRG's, essentially based on the Cayley tree:

$$N_\ell = 1, c, c(c-1), c(c-1)^2, \ldots$$

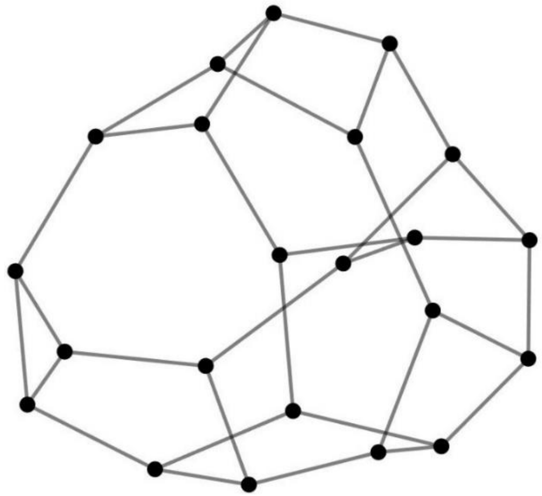
Meaning that there is an exponential growth in the volume (hyperbolic geometry)

Infinite RRGs ($N \to \infty$)



Tree graphs ($\mathrm{c} = 3$)

Finite RRGs ($N < \infty$)



Graphs with cycles ($\mathrm{c} = 3$)

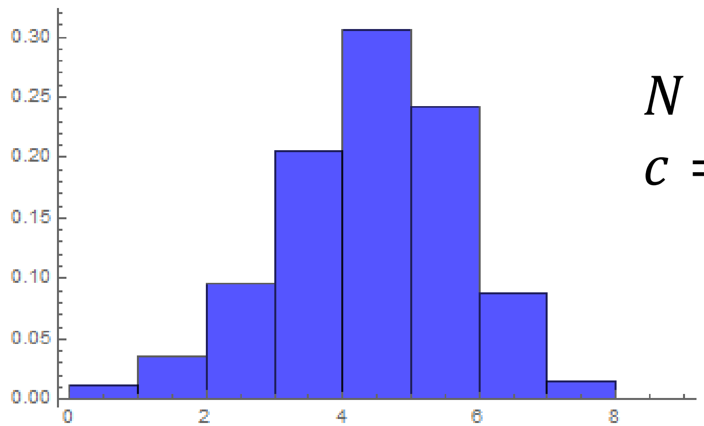
The tail distribution

It is convenient to study the tail-distribution (also known as the survival probability) defined by $P_N(L > \ell) = \sum_{\ell' > \ell} P_N(L = \ell')$

For example, $P_N(L > 1) = 1 - p$

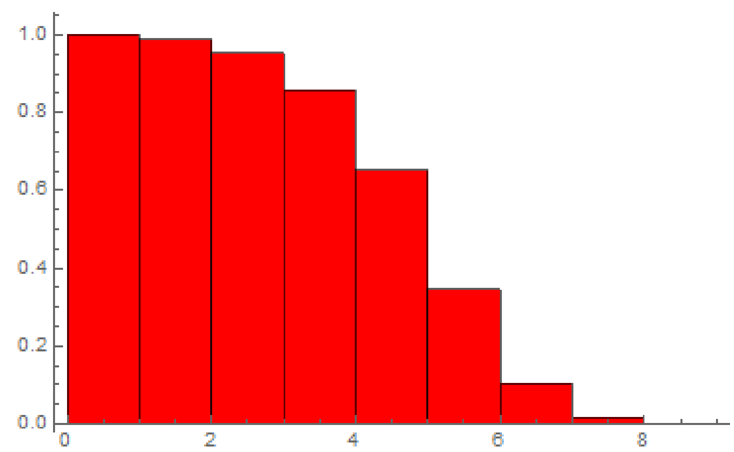
Last, $P_N(L = \ell) = P_N(L > \ell - 1) - P_N(L > \ell)$

PDF - $P_N(L = \ell)$



$N = 100$
$c = 3$

The tail-distribution
$P_N(L > \ell)$



getting started – $\boldsymbol{ER(N,p)}$...

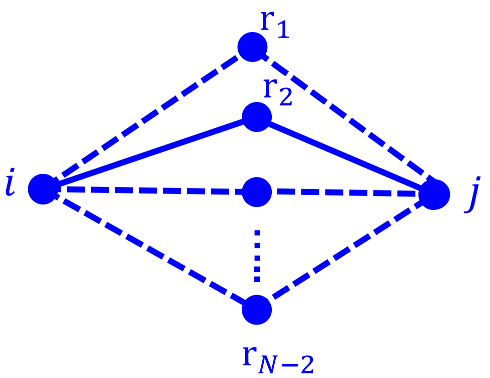
$$P_N(L > 1) = 1 - p$$

$$P_N(L > 2) = ?$$

getting started – $ER(N,p)$…

$P_N(L > 1) = 1 - p$

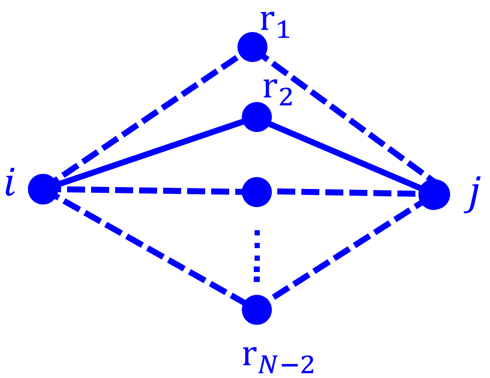
$P_N(L > 2) = (1 - p) \times ?$



getting started – $ER(N,p)$…

$$P_N(L>1)=1-p$$

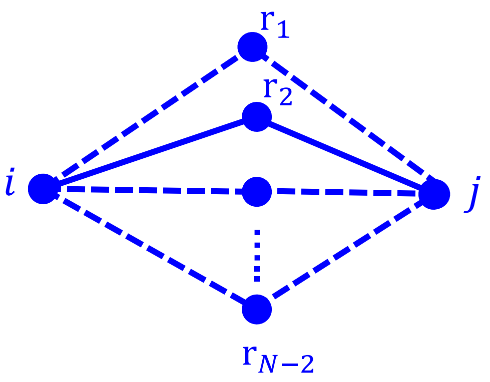
$$P_N(L>2)=(1-p)\times(1-p^2)^{N-2}$$ **Exact!**



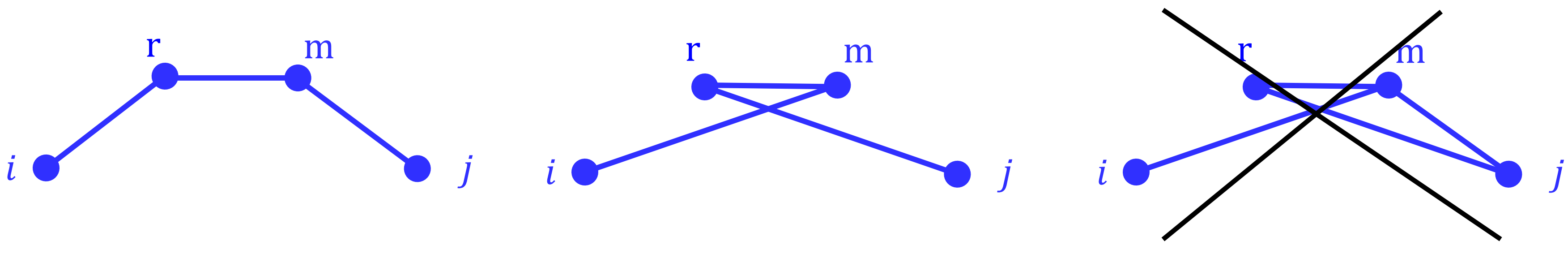
getting started – $ER(N,p)$…

$P_N(L>1) = 1-p$

$P_N(L>2) = (1-p)\times(1-p^2)^{N-2}$ **Exact!**



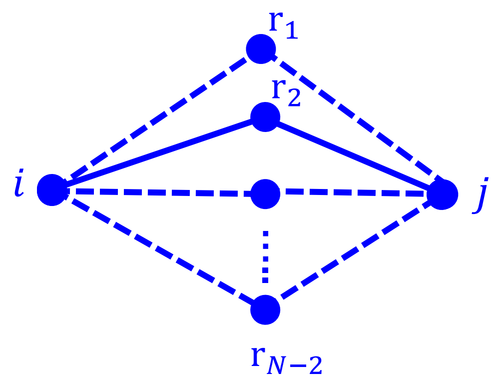
$P_N(L>3) = (1-p)\times(1-p^2)^{N-2}\times ?$



getting started – $ER(N, p)$…

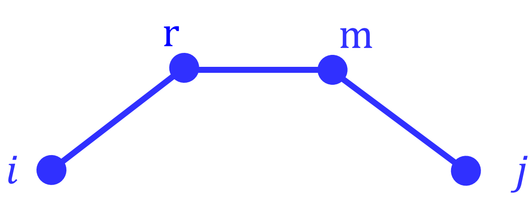
$P_N(L > 1) = 1 - p$

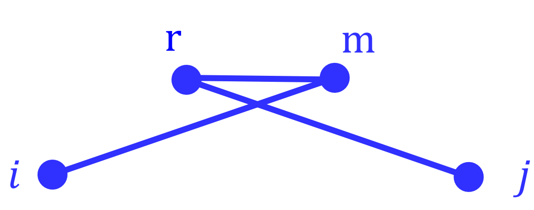
$P_N(L > 2) = (1 - p) \times \underline{(1 - p^2)^{N-2}} = P_N(L > 1) \times \underline{P_N(L > 2|L > 1)}$

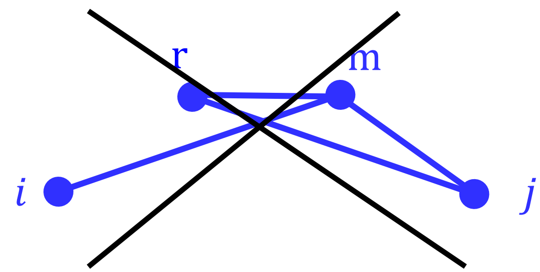


$P_N(L > 3) = P_N(L > 2) \times P_N(L > 3|L > 2)$

And the questions is $P_N(\underline{L} > 3|\underline{L} > 2) = ?$

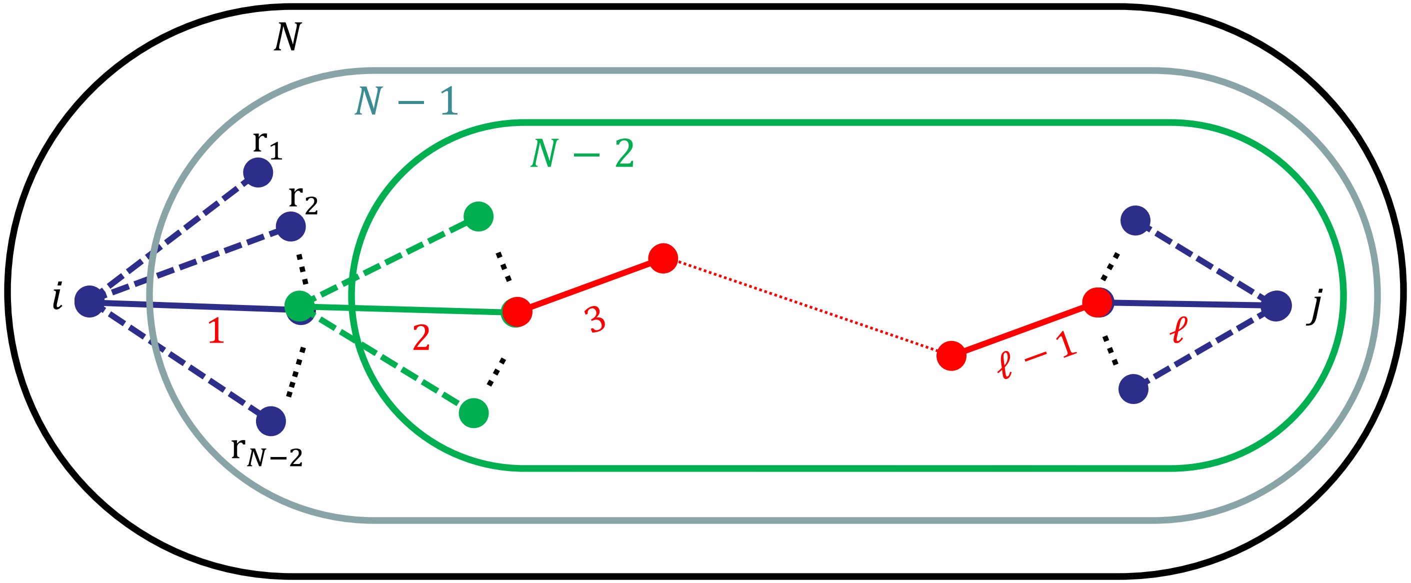






Recursive Paths Approach (RPA)

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, thus decimating our degrees of freedom (RG):



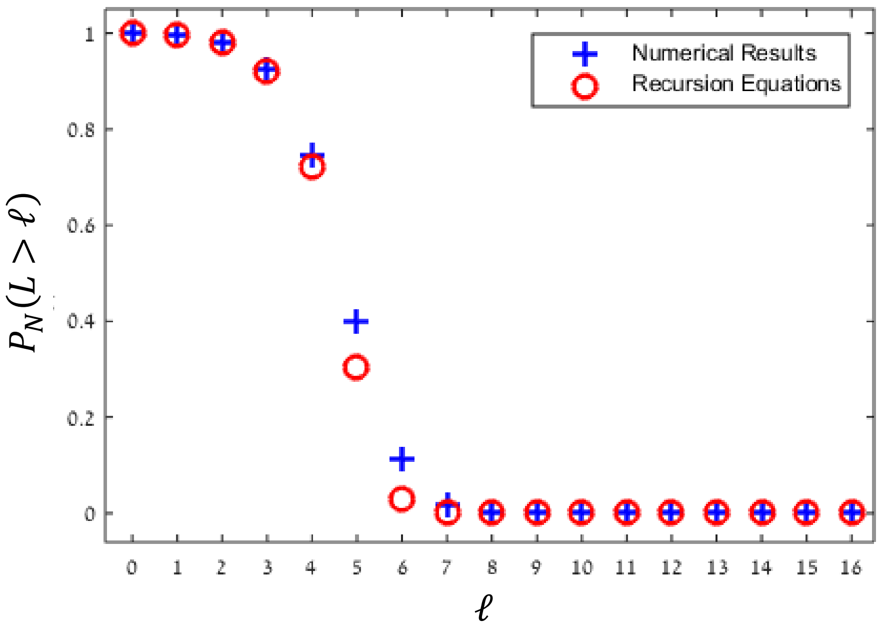
$$P_N(L>\ell|L>\ell-1) = [(1-p)+pP_{N-1}(L>\ell-1|L>\ell-2)]^{N-2}$$

with $P_{N'}(L>1|L>0) = P_{N'}(L>1) = 1-p$

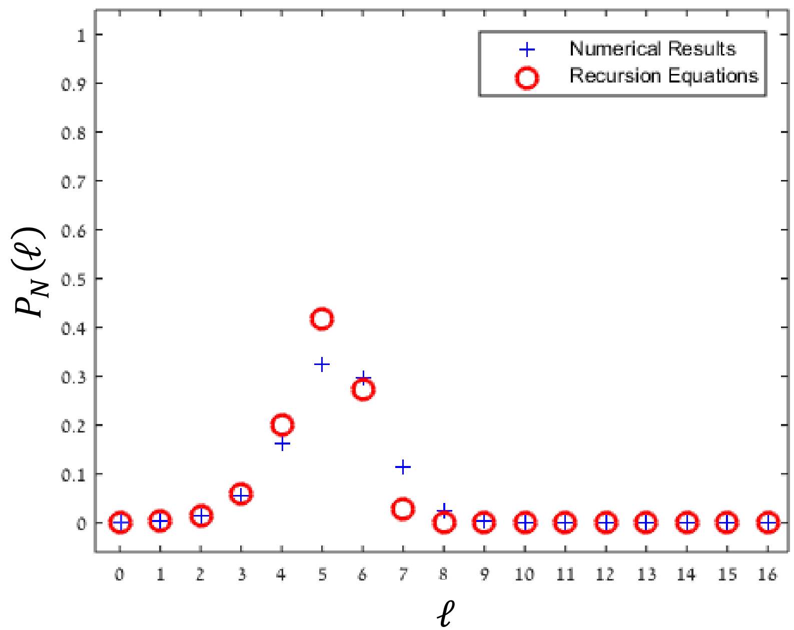
RPA – ER – results

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

The tail distribution



The PDF



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

The agreement improves with larger N.

The Configuration Model

Generalizing these results to the configuration model, two issues arise:

1. Moving away from the edge independence of the ER network to constraining $p(k)$, which requires formulating the recursion equations using the degree distribution rather than relying on independence. This can be done, if only as an exercise, for the ER case by imposing a Poissonian degree distribution.
2. Choosing a random node in a network indeed samples uniformly from the degree distribution $p(k)$. However, random neighbor of a random node is typically of a higher degree. Actually, this sampling is proportional to the degree, and hence the higher the degree, the more probable it is to be choses, $\tilde{p}(k) = \frac{k}{<K>} p(k)$ (cavity method...). This effect has no ER counterpart.

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

One can write down the recursion equations for these quantities

$$\chi\left(L_{ij} > \ell | L_{ij} > \ell - 1\right) = \prod_{\mathrm{r} \neq i,j} \left[1 - A_{i\mathrm{r}} + A_{i\mathrm{r}} \chi\left(L_{\mathrm{r}j} > \ell - 1 | L_{\mathrm{r}j} > \ell - 2\right)\right]$$

and average these equations over the ensemble to obtain the RPA equations for the Erdős-Rényi case. This can be achieved using

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

RPA and the Cavity Method

Using

$$\chi\left(L_{ij} > \ell | L_{ij} > \ell - 1\right) = \prod_{\mathrm{r}\neq i,j} \left[1 - A_{i\mathrm{r}} + A_{i\mathrm{r}}\chi\left(L_{\mathrm{r}j} > \ell - 1 | L_{\mathrm{r}j} > \ell - 2\right)\right]$$

$$P(A) = \prod_{i<j} \left[(1-p)\delta_{A_{ij},0} + p\delta_{A_{ij},1}\delta_{A_{ji},1}\right]$$

$$P_N(L > \ell | L > \ell - 1) = \langle \chi\left(L_{ij} > \ell | L_{ij} > \ell - 1\right)\rangle$$

and averaging over the ensemble to obtain the RPA equations for the Erdős-Rényi case, we recover

$$P_N(L > \ell | L > \ell - 1) = [(1-p) + pP_{N-1}(L > \ell - 1 | L > \ell - 2)]^{N-2}$$

with the initial condition

$$P_{N'}(L > 1 | L > 0) = P_{N'}(L > 1) = 1 - p$$

RPA for the Configuration model

Using the Cavity formalism one can easily generalize this approach to deal with the configuration model.

The starting point are again the indicator functions $\chi(L_{ij} > \ell | L_{ij} > \ell - 1)$

One can write down the recursion equations for these quantities

$$\chi(L_{ij} > \ell | L_{ij} > \ell - 1) = \prod_{\mathrm{r} \neq i,j} \left[1 - A_{i\mathrm{r}} + A_{i\mathrm{r}} \chi^{(i)}(L_{\mathrm{r}j} > \ell - 1 | L_{\mathrm{r}j} > \ell - 2)\right]$$

$$= \prod_{\mathrm{r} \in \partial_i \setminus j} \left[\chi^{(i)}(L_{\mathrm{r}j} > \ell - 1 | L_{\mathrm{r}j} > \ell - 2)\right]$$

where $\chi^{(i)}(L_{\mathrm{r}j} > \ell - 1 | L_{\mathrm{r}j} > \ell - 2)$ are the cavity indicator functions, i.e. on a graph which does not contain i, obeying a similar equation.

$$\chi^{(i)}(L_{\mathrm{r}j} > \ell | L_{\mathrm{r}j} > \ell - 1) = \prod_{s \in \partial_r \setminus j} \left[\chi^{(r)}(L_{sj} > \ell - 1 | L_{sj} > \ell - 2)\right]$$

RPA for the Configuration model

Using the general degree distribution $p(k)$ to average over these equations (assuming a *local* tree-like structure), the general equations become

$$P_N(L > \ell | L > \ell - 1) = \sum_{k=1}^{N-2} p(k) [\tilde{P}_{N-1}(L > \ell - 1 | L > \ell - 2)]^k$$

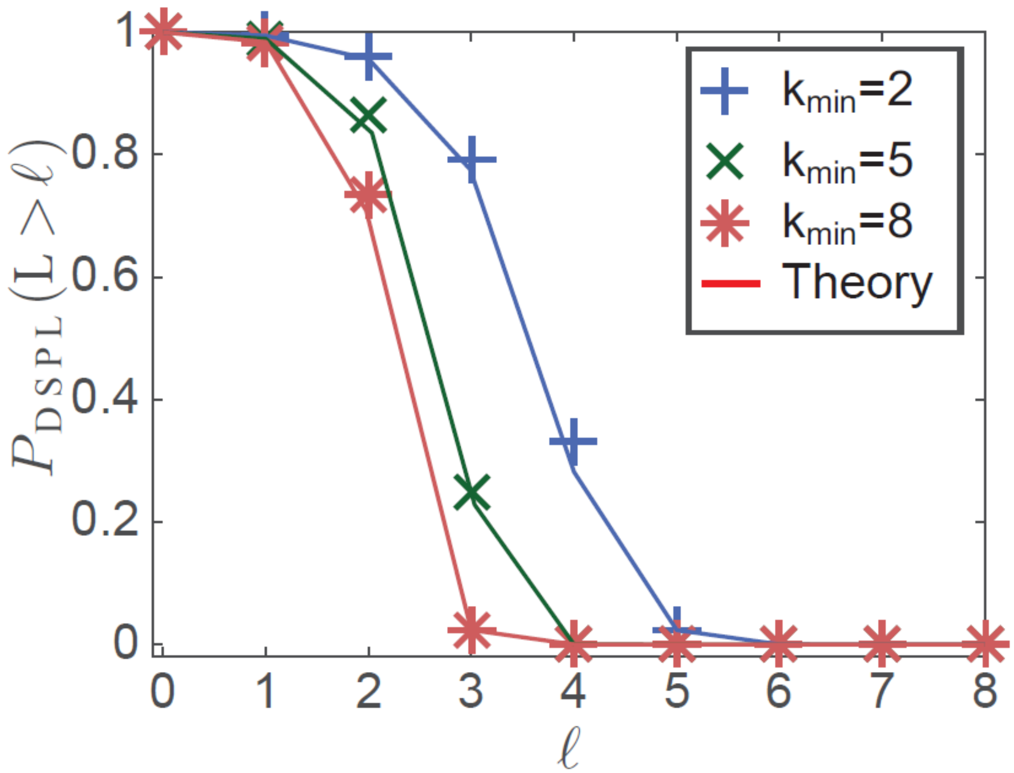
$$\tilde{P}_N(L > \ell | L > \ell - 1) = \sum_{k=1}^{N-2} \frac{k}{c} p(k) [\tilde{P}_{N-1}(L > \ell - 1 | L > \ell - 2)]^{k-1}$$

with the initial conditions

$$P_N(L > 1 | L > 0) = \sum_{k=1}^{N-2} p(k) \left(1 - \frac{1}{N-1}\right)^k \simeq 1 - \frac{\langle k \rangle}{N-1}$$

$$\tilde{P}_N(L > 1 | L > 0) = \sum_{k=1}^{N-2} \frac{k}{c} p(k) \left(1 - \frac{1}{N-1}\right)^{k-1} \simeq 1 - \frac{\langle k^2 \rangle / \langle k \rangle - 1}{N-1}$$

DSPL using the cavity method – results: scale free



$N = 10^3$

Scale-Free graph: $\mathrm{P}(k) \propto k^{-2.5}$ for $k = k_{min}, k_{min}+1, \ldots, k_{max}$

RPA for the Configuration model

We can actually study the distribution over the ensemble using

$$\pi_\ell(P_N(L > \ell | L > \ell - 1) = m) = \sum_{k=1}^{N-2} \frac{k}{c} p(k) \int_0^1 \int_0^1 \cdots \int_0^1 \prod_{\nu=1}^{k} \pi_{\ell-1}(m_\nu) dm_\nu \delta\left(m - \prod_{\nu=1}^{k} m_\nu\right)$$

…

This is very Reimer…

And there is one exact result…

RPA for Random Regular Graphs

Denoting

$$P(L > \ell | L > \ell - 1) = \langle \chi\big(L_{ij} > \ell | L_{ij} > \ell - 1\big) \rangle$$

$$\tilde{P}(L > \ell | L > \ell - 1) = \langle \chi^{(i)}\big(L_{ij} > \ell | L_{ij} > \ell - 1\big) \rangle$$

we get the following recursion equations

$$P(L > \ell | L > \ell - 1) = [\tilde{P}_{N-1}(L > \ell - 1 | L > \ell - 2)]^{c}$$

$$\tilde{P}(L > \ell | L > \ell - 1) = [\tilde{P}_{N-1}(L > \ell - 1 | L > \ell - 2)]^{c-1}$$

with the initial conditions

$$P(L > 1 | L > 0) = \left(1 - \frac{1}{N-1}\right)^{c} \simeq 1 - \frac{c}{N-1}$$

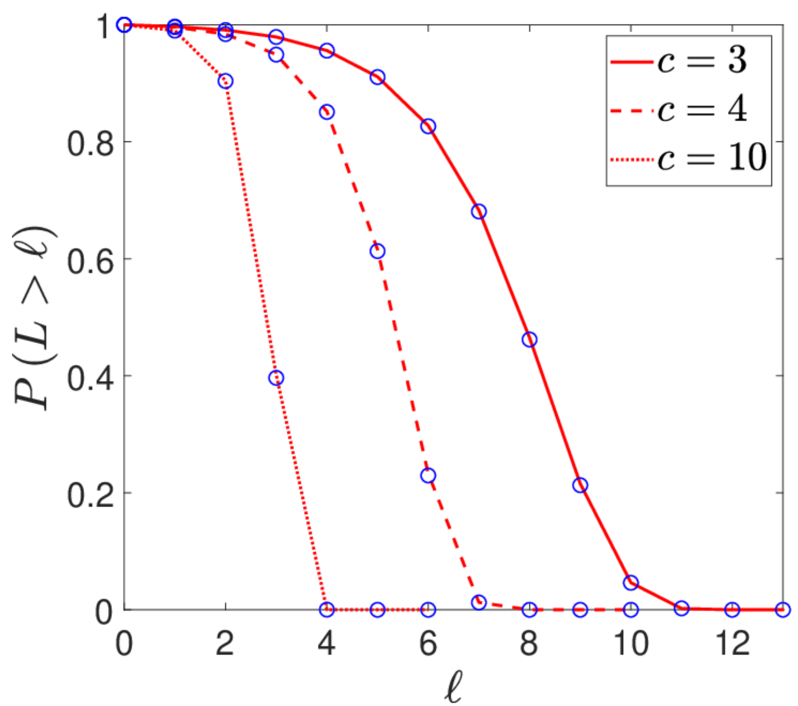
$$\tilde{P}(L > 1 | L > 0) = \left(1 - \frac{1}{N-1}\right)^{c-1} \simeq 1 - \frac{c-1}{N-1}$$

RRG: Exact result

So that eventually we get

$$P(L > \ell) = \exp\left\{-\frac{c}{(c-2)N}\left[(c-1)^{\ell}-1\right]\right\}$$

which is a discrete Gompertz distribution, encountered in survival analysis (this observation is actually quite insightful) - which is an exact result.



Digression: Gompertz distribution

The Gompertz distribution is often used in survival analysis, which deals with the study of time until an event of interest (e.g., death, failure, or disease occurrence) happens. It is particularly suitable for modeling mortality rates in human populations or the failure times of mechanical components.

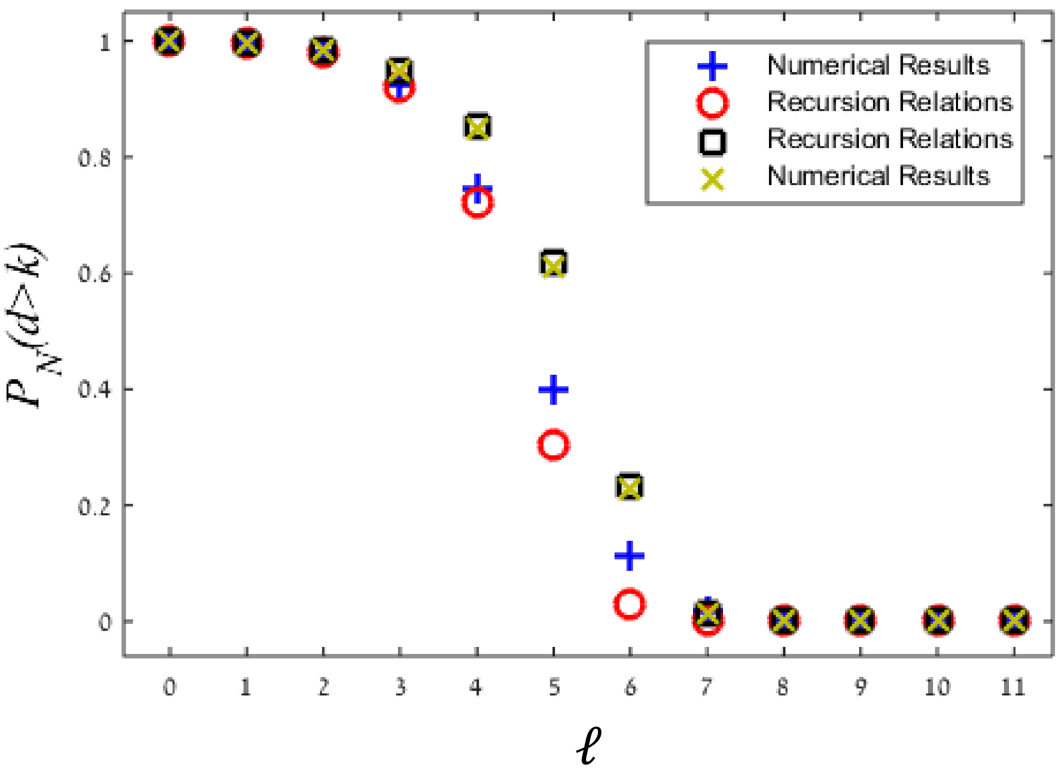
One of the defining features of the Gompertz distribution is that it exhibits an increasing hazard rate over time. This means that the likelihood of an event occurring (e.g., death or failure) becomes higher as time progresses.

Why are there quite some 100 years old people while the oldest known people are not older than 120?

(with an exponential distribution you would see many 200 years old).

RRG vs ER

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



Erdős-Rényi:

$$\langle L \rangle = \frac{\ln N}{\ln c}$$

Random Regular Graph:

$$\langle L \rangle = \frac{\ln N}{\ln(c-1)}$$

Also note that distances in the ER graph are slightly shorter than in a regular graph with the same mean degree – not obvious.

RRG: Mean and Variance

We got interested in the mean and variance of the RRG. The mean was known to leading order:

$$\langle L \rangle = \frac{\ln N}{\ln(c-1)} + O(1),$$

while the variance was not known at all. This seemed like a small exercise, but it turned out that there are many surprises.

In short: having an explicit expression for the distribution does not imply that the mean is trivial…

The tail-sum formula says that

$$\langle L \rangle = \sum_{\ell} P(L > \ell) = \sum_{\ell} \exp\left\{-\frac{c}{(c-2)N}\left[(c-1)^{\ell} - 1\right]\right\}$$

However, this is not a easy to evaluate even asymptotically. More surprises follow.

RRG: Mean and Variance

In a recent paper [Tishby, Biham, Kühn and Katzav, *J. Phys. A* **55**, 265005 (2022)] we focused on calculating the mean and variance of this distribution. Actually we succeeded calculating the (discrete) Laplace transform

$$\mathcal{L}\{P(L>\ell)\}(s) = \sum_{\ell=0}^{N-2} \mathrm{e}^{-s\ell} P(L>\ell)$$

using the Euler-Maclaurin summation

$$\begin{aligned}\mathcal{L}\{P(L>\ell)\}(s) = {} & \frac{1}{\ln(c-1)} \mathrm{E}_{1+\frac{s}{\ln(c-1)}} \left[\left(\frac{c}{c-2} \right) \frac{1}{N} \right] \\ & + \frac{1}{2} + \left[\frac{1}{2} \coth\left(\frac{s}{2} \right) - \frac{1}{s} \right] + \mathcal{O}\left(\frac{1}{N} \right)\end{aligned}$$

From this we can easily obtain the moment generating function, and hence the mean and variance.

RRG: Mean and Variance

The mean Distance:

$$\boxed{\langle L\rangle = \frac{\ln N}{\ln(c-1)}} + \frac{1}{2} - \frac{\ln\left(\frac{c}{c-2}\right)+\gamma}{\ln(c-1)} + O\left(\frac{\ln N}{N}\right)$$

(improving over the known leading order).

This differs from the mean Diameter [Bollobas 1982]

$$\langle D\rangle = \frac{\ln N}{\ln(c-1)} + \frac{\ln\ln N}{\ln(c-1)} + O(1)$$

And the variance is

$$\mathrm{Var}(L) = \frac{\pi^2}{6[\ln(c-1)]^2} + \frac{1}{12} + O\left(\frac{1}{N}\right)$$

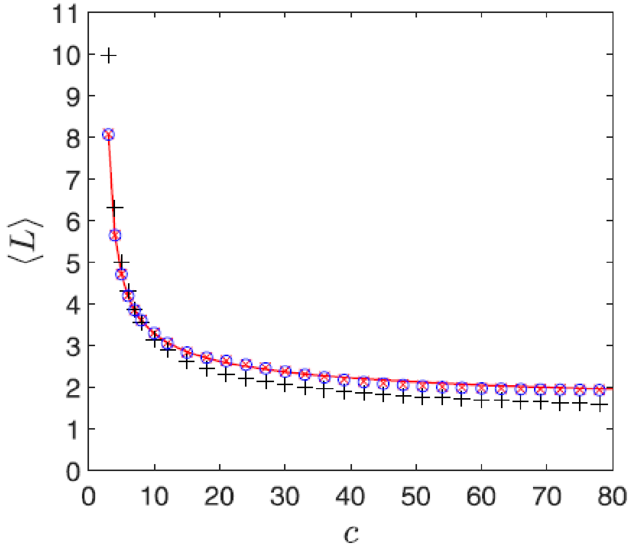
RRG: Mean and Variance

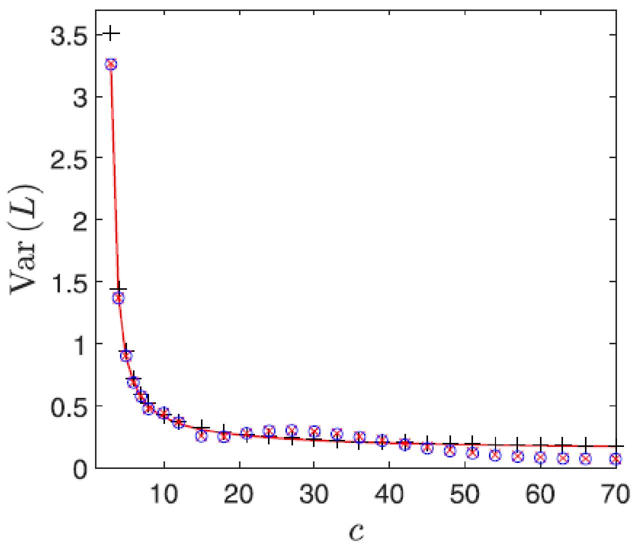
The mean Distance:

$$\boxed{\langle L\rangle = \frac{\ln N}{\ln(c-1)}} + \frac{1}{2} - \frac{\ln\left(\frac{c}{c-2}\right) + \gamma}{\ln(c-1)} + O\left(\frac{\ln N}{N}\right)$$

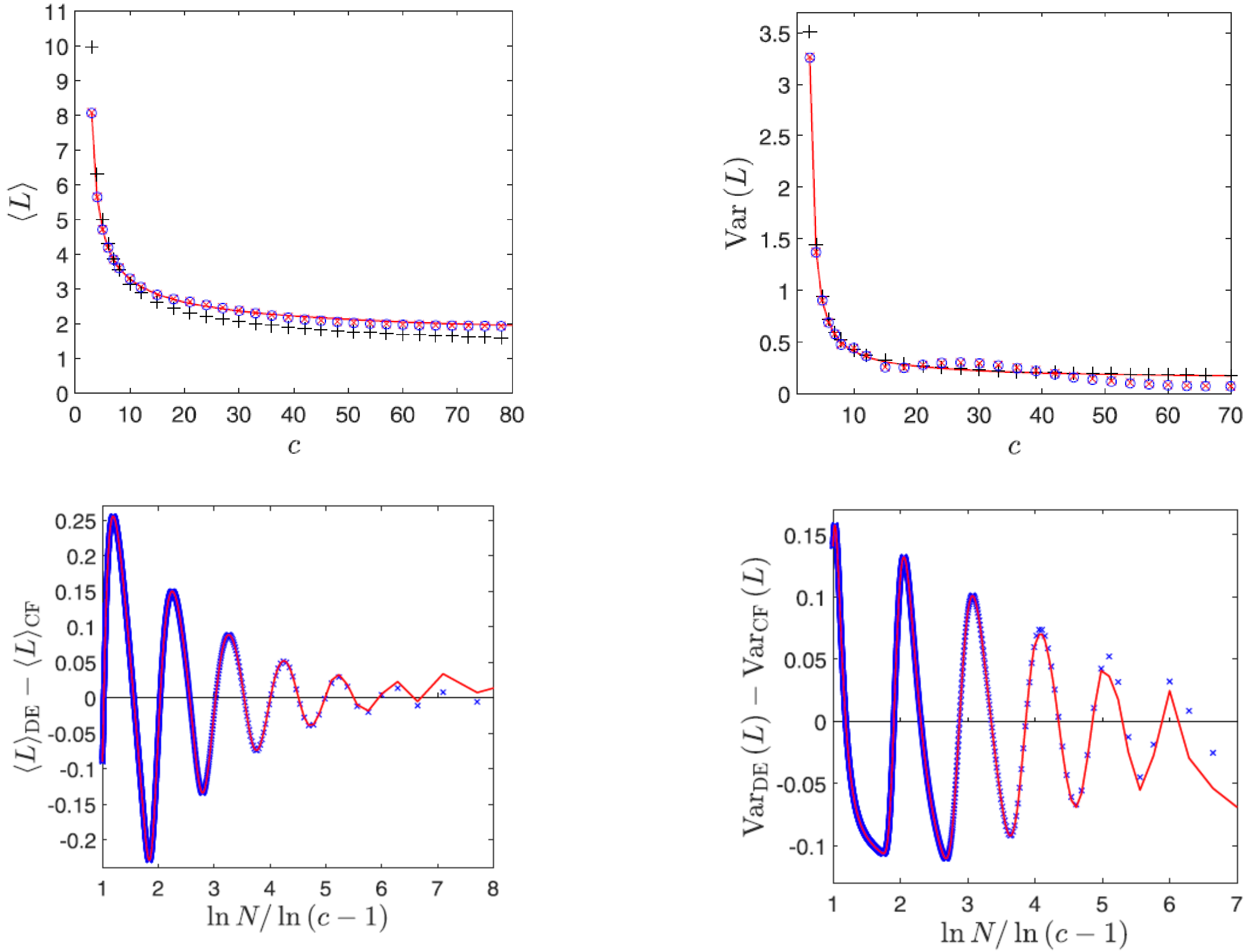
(improving over the known leading order), and the variance is

$$\mathrm{Var}(L) = \frac{\pi^2}{6[\ln(c-1)]^2} + \frac{1}{12} + O\left(\frac{1}{N}\right)$$

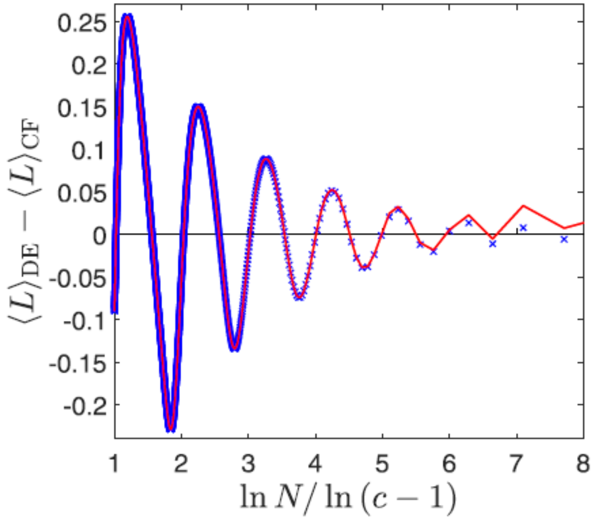


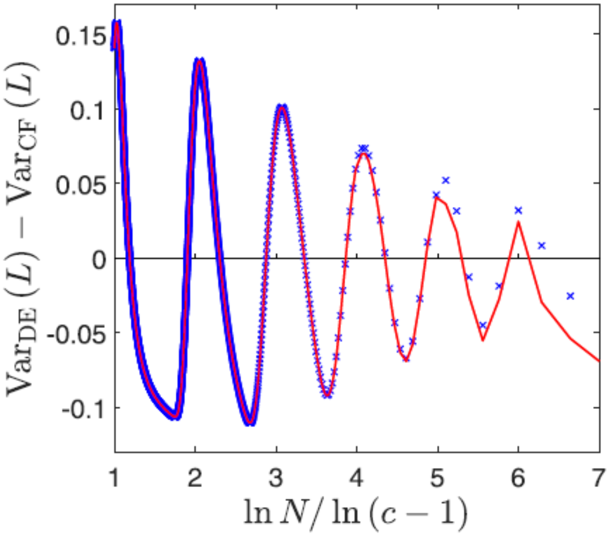


Mean and Variance: oscillations



Mean and Variance: oscillations





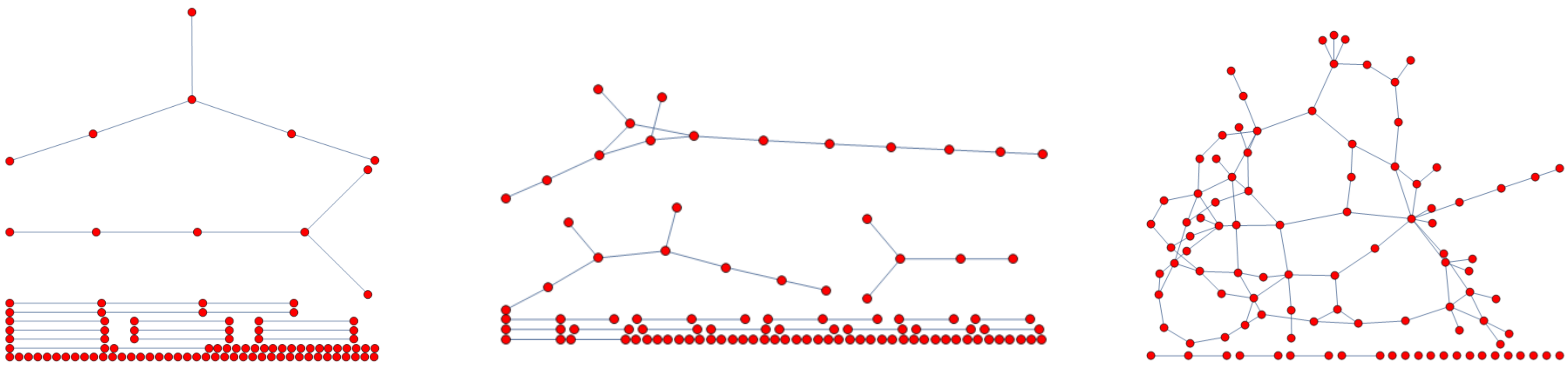
We could show that these differences exhibit oscillations as a function of $\frac{\ln N}{\ln(c-1)}$ with period 1. We also devised efficient simple approximations (solid red lines) that recover these oscillations well, except for small values of c.

In essence, it is related to the discreteness

$$\phi = \frac{\ln N}{\ln(c-1)} - \left\lfloor \frac{\ln N}{\ln(c-1)} \right\rfloor$$

Focus - reminder

In this talk we focused on ER networks in the super-percolating regime. However below we will briefly discuss the subpercolating regime too $c < 1$.

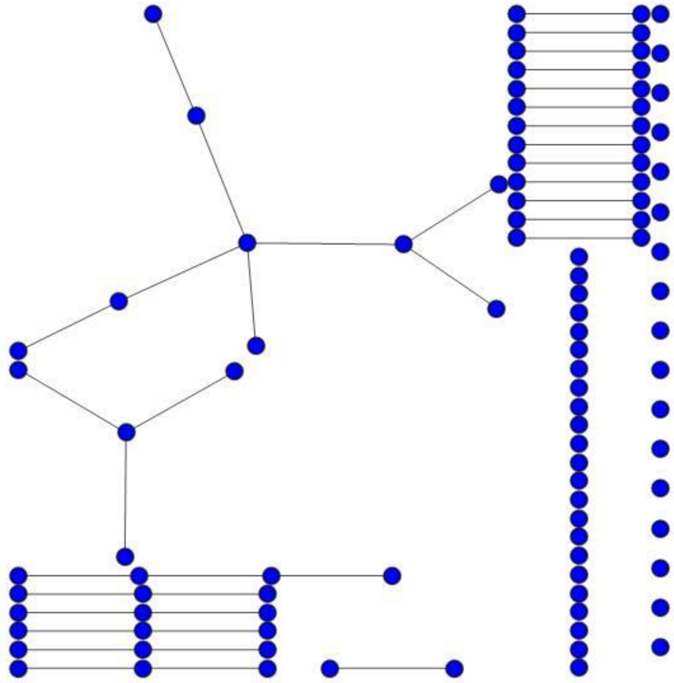


$\mathrm{N} = 100, c = 0.5$ $\mathrm{N} = 100, c = 1$ $\mathrm{N} = 100, c = 2$

Subcritical case – $\mathrm{ER}(\mathrm{N},\mathrm{p}),\ c < 1$

$\mathrm{N} = 100, c = 0.7$, namely below the percolation threshold at $c = 1$ (sub-critical). It consists of isolated trees (forest).

First, the probability that two nodes are connected is fairly low, namely $P(L < \infty) = \frac{c}{(1-c)N}$



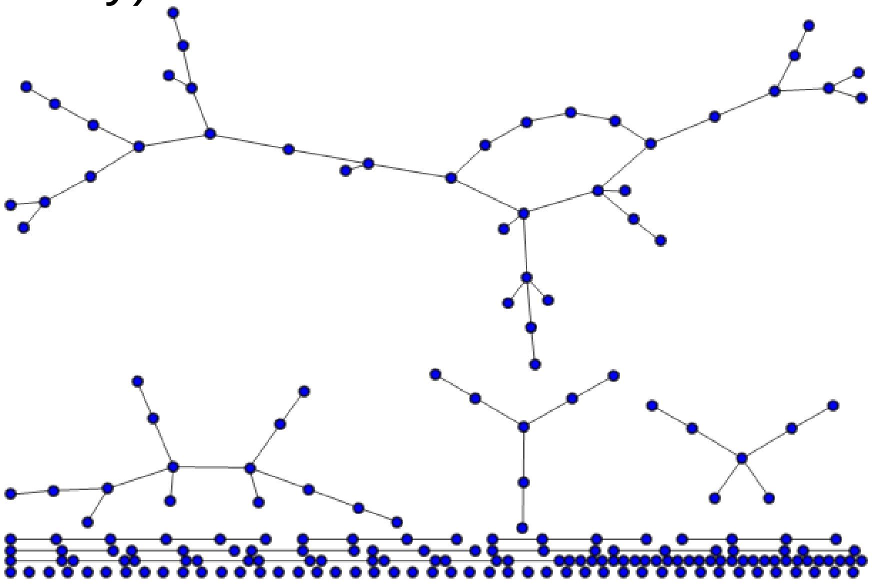
$\mathrm{N} = 100,\ \mathrm{c} = 0.7$

Katzav, Biham & Hartman, PRE **98** 012301 (2018)

Subcritical case – *Why?*

There are several motivations to understand this case, beyond basic intellectual curiosity.

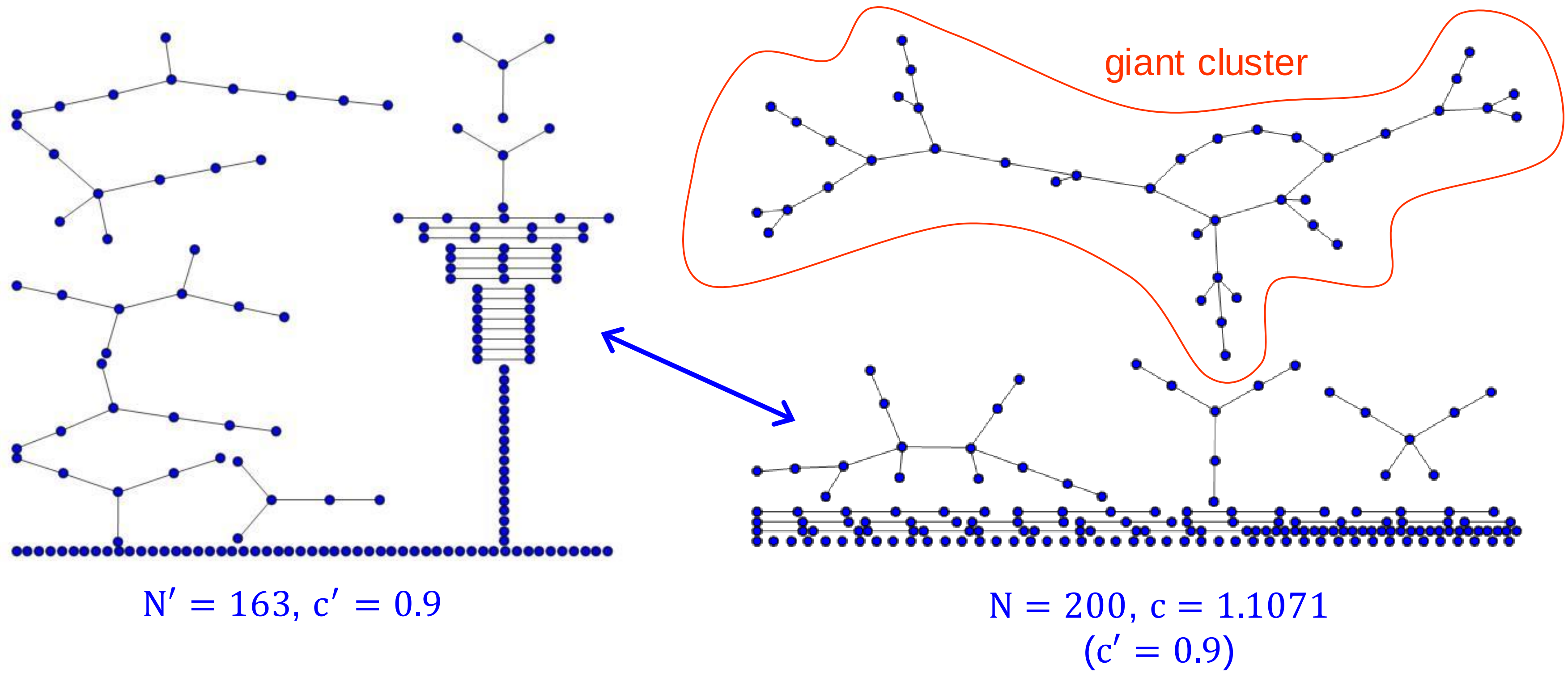
1. Even above the percolation transition there are small components, especially near/at the transition (Duality)
2. The tree branches that stick out from the 2-core are statistically the same as the small tree components (switching).
3. The resistance distance on trees is identical to the Shortest Path Length – since the shortest path is the only path.



$N = 200,\ c = 1.1$

Subcritical case – *Duality*

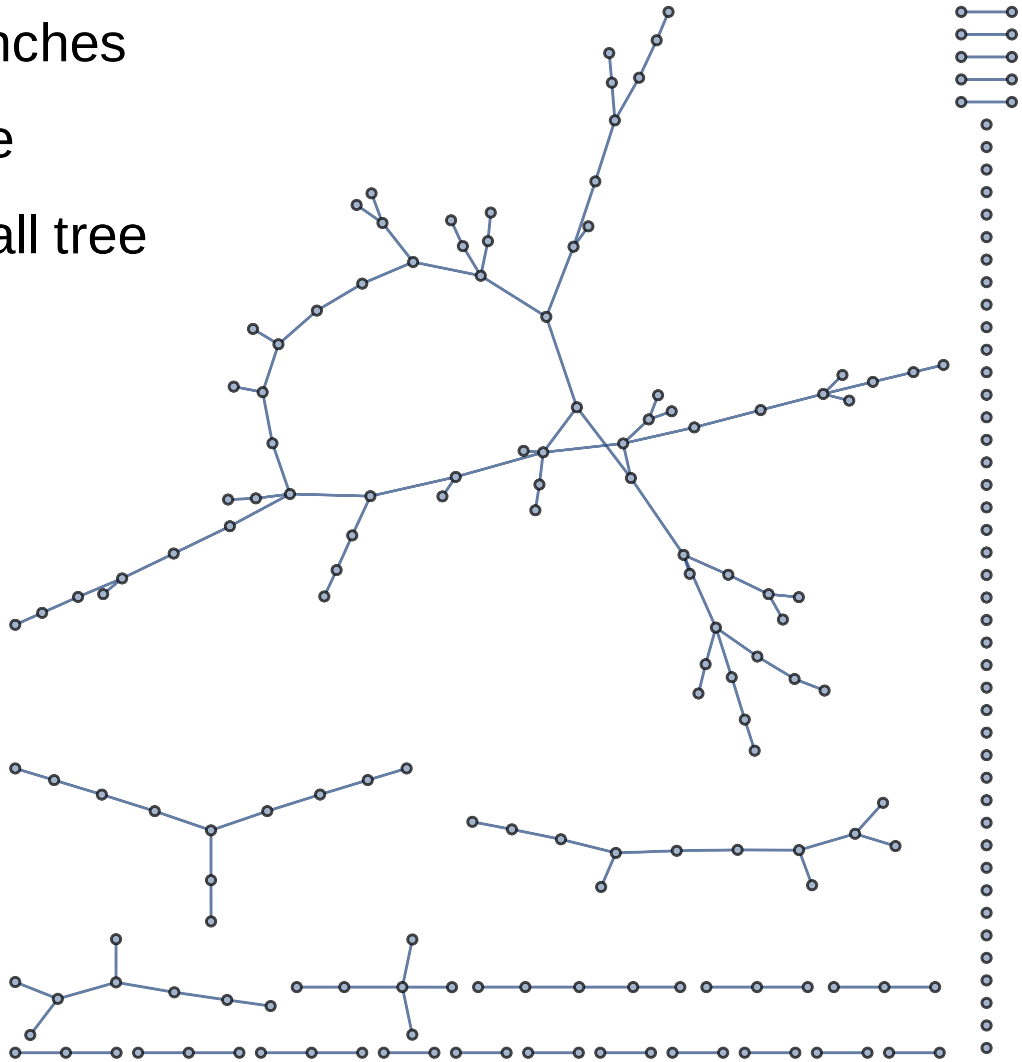
In ER networks there is a (known) duality between finite components above the percolation and a corresponding sub-percolating network $ER(N, c > 1) \rightarrow ER(N', c' < 1)$, where $N' = N(1-g)$, and $c' = c(1-g)$



Katzav, Biham & Hartman, PRE **98** 012301 (2018)
I. Tishby, O. Biham, E. Katzav and R. Kühn (2018)

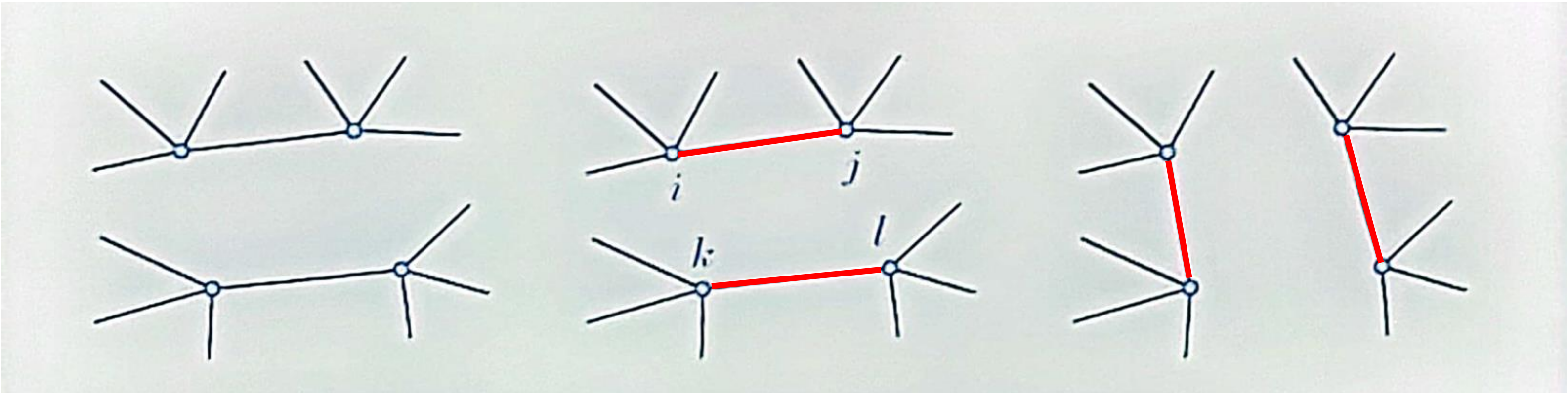
Tree branches – switching

The tree branches that tree branches that stick out from the 2-core are statistically the same as the small tree components (switching).



$N = 200,\ c = 1.4$

Equilibration: switching



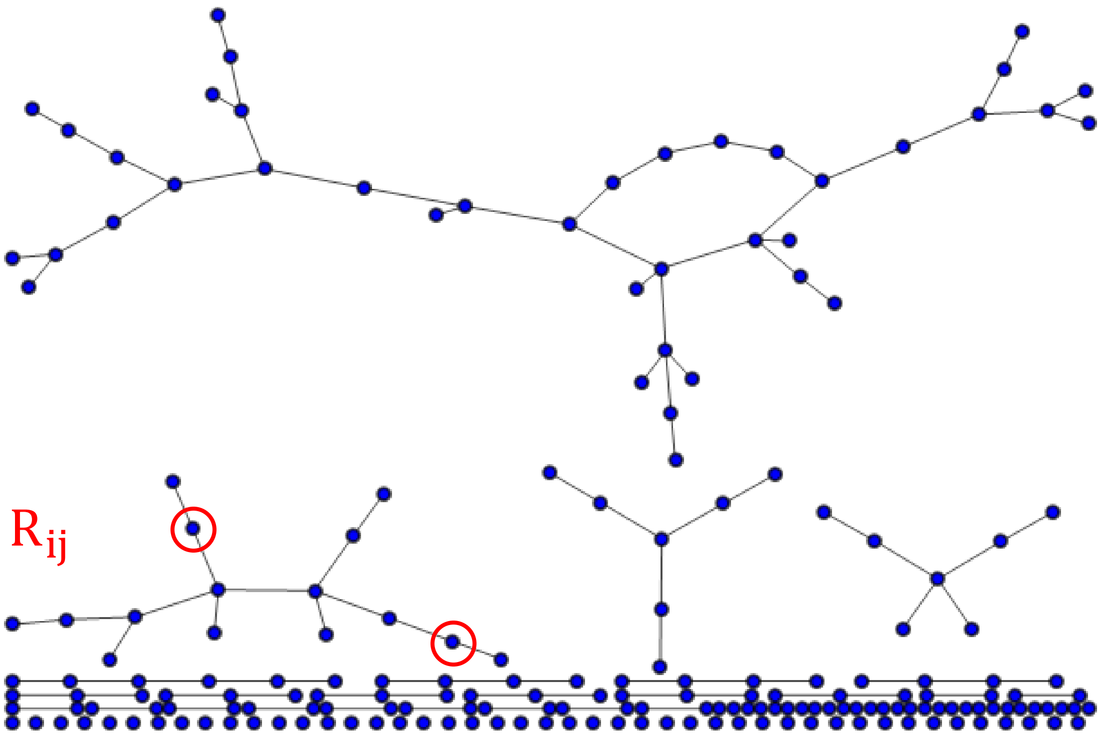
Note that switching does not alter the degree distribution!

Configuration network ensembles are invariant with respect to such transformations. In fact such steps are “equilibration” transformations, in the sense that if you start with any network it converges onto the uncorrelated configuration network ensemble.

Subcritical case – *Resistance*

The resistance distance on trees is identical to the Shortest Path Length – since the shortest path is the only path.

$L_{ij} = 4 = R_{ij}$



$N = 200,\ c = 1.1$

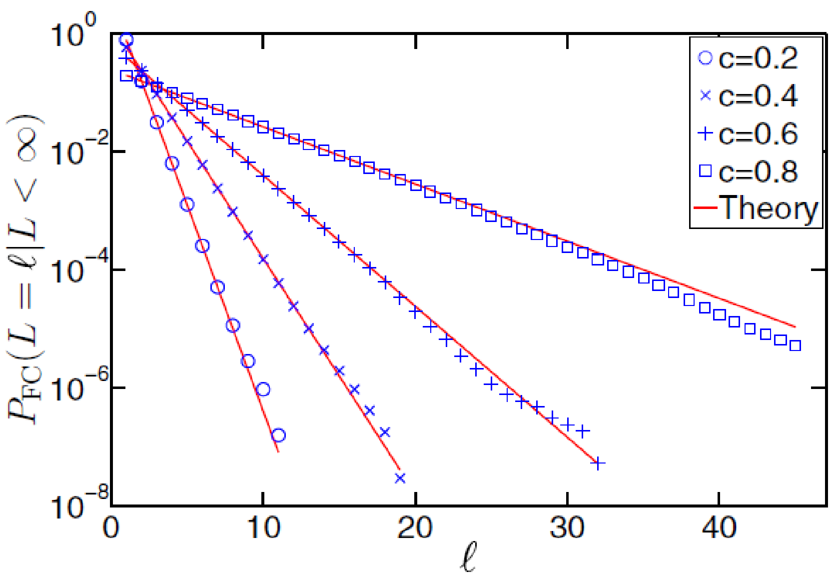
Subpercolating ER: Exact result…

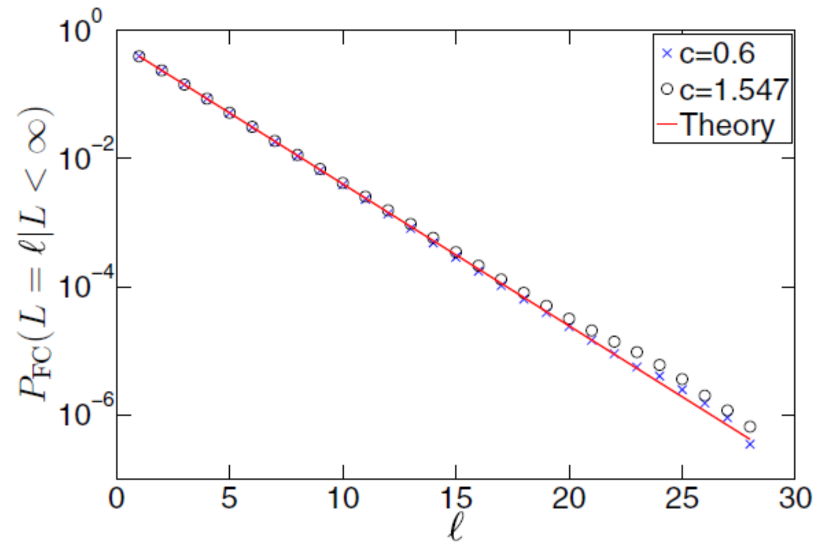
This suggested the following geometric (exponential) distribution

$$P(L=\ell|L<\infty)=P(L=\ell|L<\infty,S\leq s\to\infty)=(1-c)c^{\ell-1}$$

which could be argued using a mean-field argument, regarding the expansion of shells around a random node: $1,c,c^2,c^3,\ldots$ (up to normalization by $(1-c)$).

However, the exact enumeration based on the topological expansion indicates that this is an exact result, leading to various conclusions



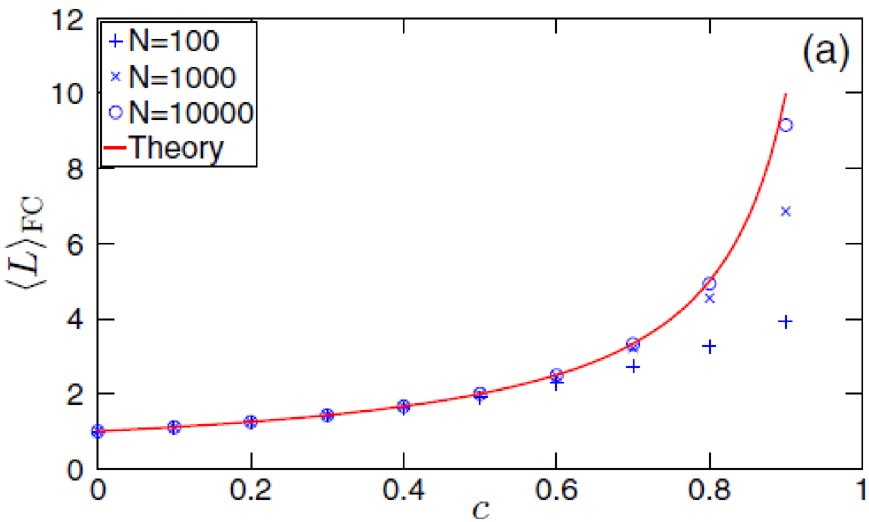


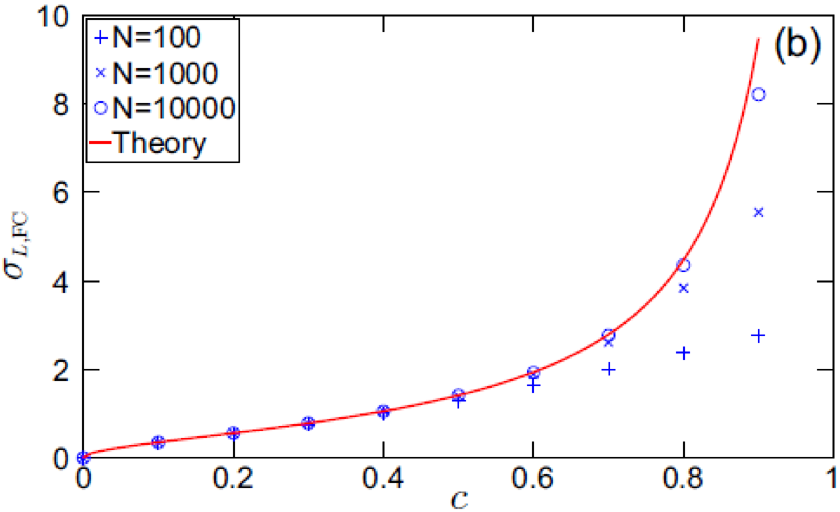
Subpercolating ER: cumulants

Hence we get

$$E[L|L < \infty] = \frac{1}{1-c}$$

$$\mathrm{Var}[L|L < \infty] = \frac{c}{(1-c)^2}$$

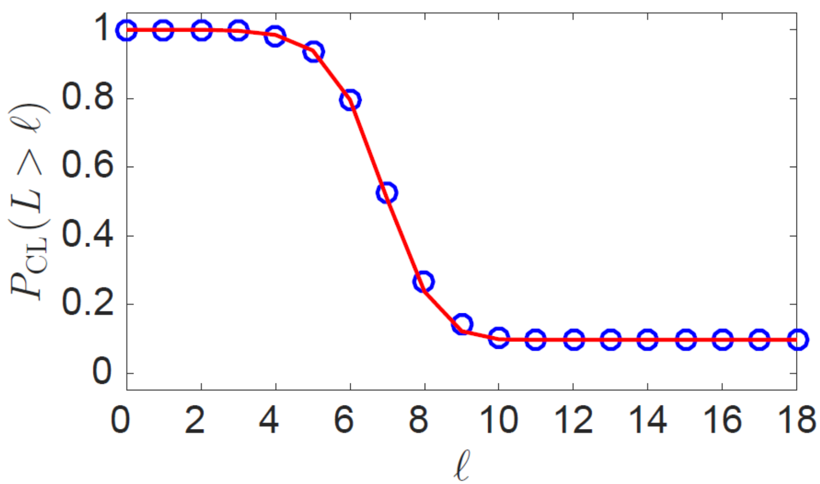


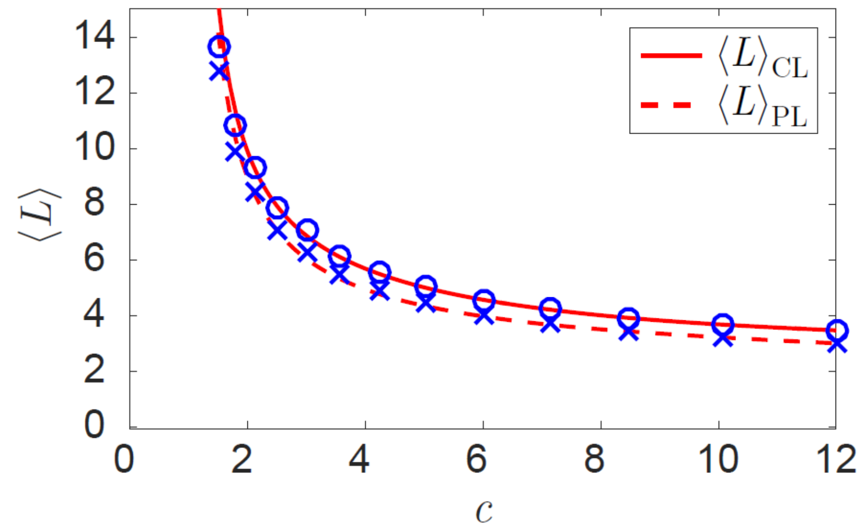


where the discrepancy is probably related to finite size effects (big trees do not appear unless N is large enough)

Shortest Cycles

$\mathrm{ER}(N = 10^4, c = 4)$





We have developed an approach which is valid **within the configuration model**.

Excellent agreement between theory (lines) and simulations (symbols)

On average, cycles are longer than shortest paths by 1.

Bonneau, Hassid, Biham, Kühn & Katzav, PRE **96** 062307 (2017).

Node-Duplication

Recently, we developed an approach to obtain the DSPL in growing networks, such as the node-duplication model.

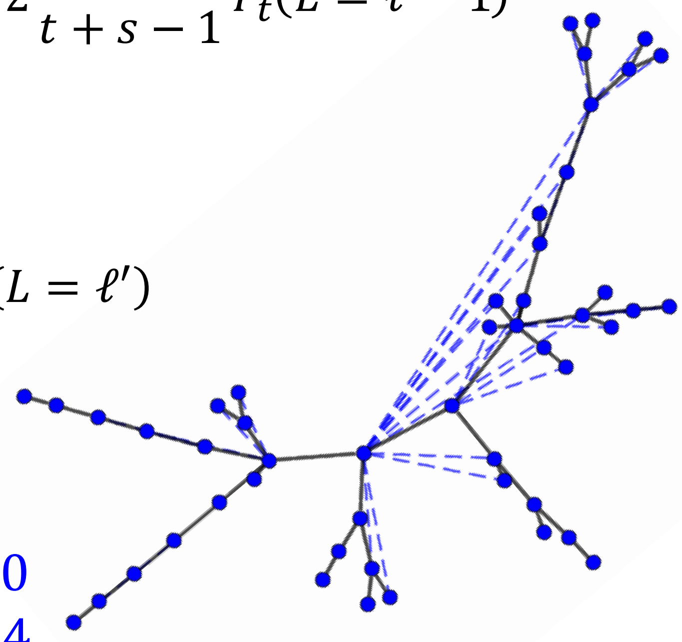
Being an out of equilibrium and growing network, the approach is based on writing a master equation where time is the network size.

$$\frac{d}{dt}P_t(L=\ell) = -2\frac{(1-\eta)(t+s)+\eta}{(t+s-1)(t+s+1)}P_t(L=\ell) + 2\frac{1-\eta}{t+s-1}P_t(L=\ell-1)$$

hence

$$P_t(L=\ell) = \left(\frac{s-1}{s+1}\right)\left(\frac{t+s+1}{t+s-1}\right)\sum_{\ell'=1}^{\min\{\ell,\Delta_0\}}\frac{e^{-c_t}c_t^{\ell-\ell'}}{(\ell-\ell')!}P_0(L=\ell') + \frac{1}{(1-\eta)(s+1)}\sum_{\ell'=0}^{\infty}\frac{e^{-c_t}c_t^{\ell+\ell'}}{(\ell+\ell')!}e^{-\mu\ell'}$$

$s=$ initial size, $\eta=$ degeneracy

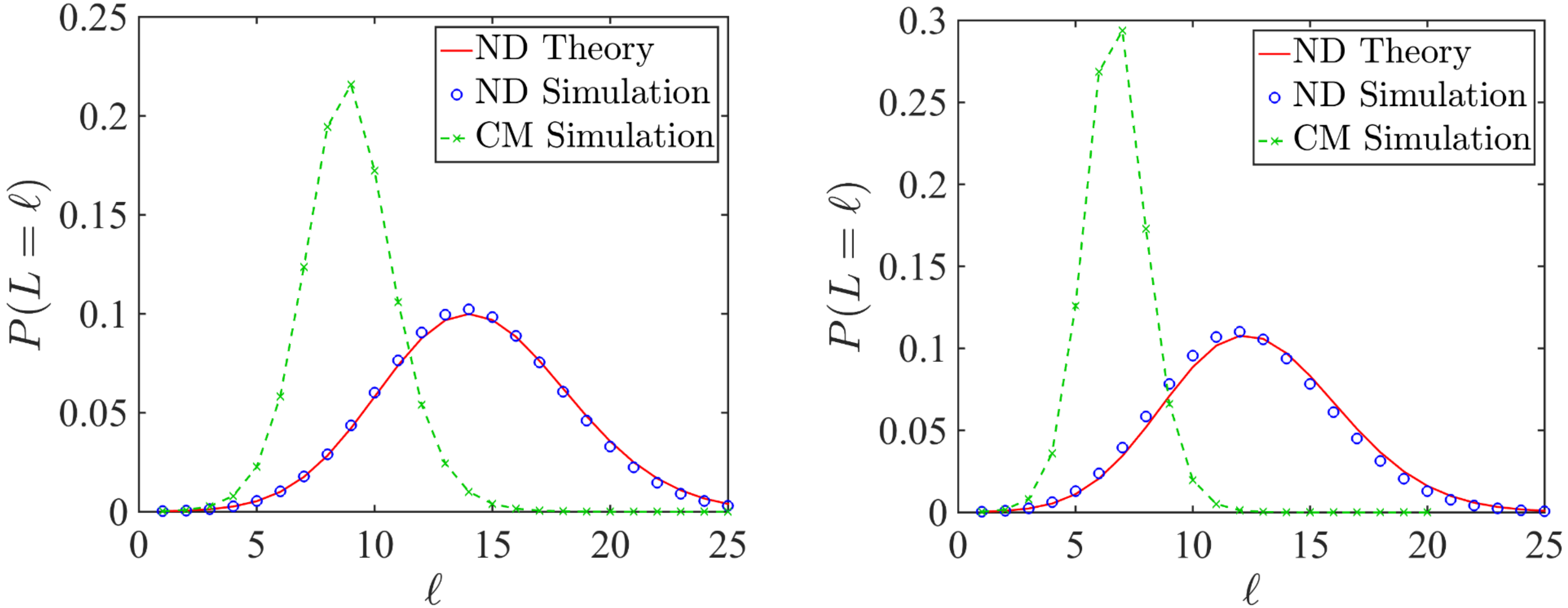


Node-Duplication Model

Exact solution: $P_t(L = \ell) = \textbf{initial} + \frac{1}{(1-\eta)(s+1)} \sum_{\ell'=0}^{\infty} \frac{e^{-c_t} c_t^{\ell+\ell'}}{(\ell+\ell')!} e^{-\mu \ell'}$

Interestingly, the DSPL is Poissonian while the degree distribution being scale-free.

$p = 0.1 \quad (\gamma = 2.86 \ldots)$ $\qquad$ $p = 0.2 \quad (\gamma = 2.65 \ldots)$



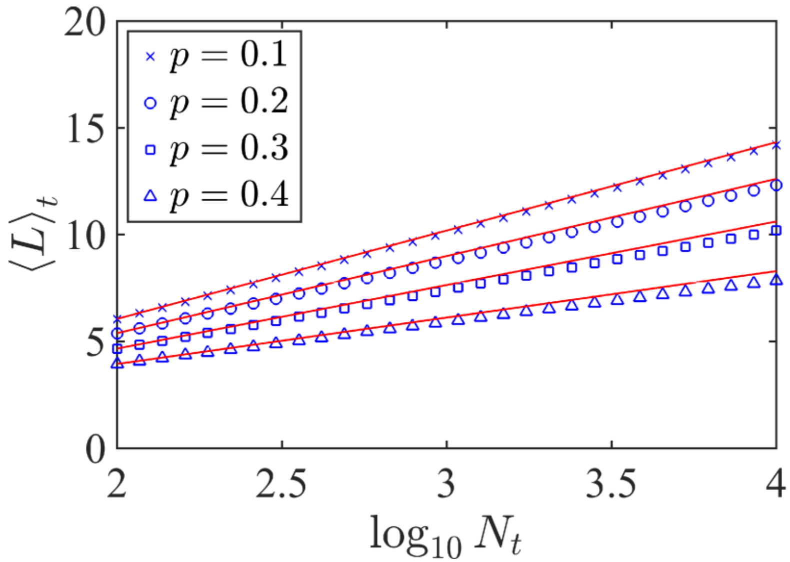
Excellent agreement with the simulation

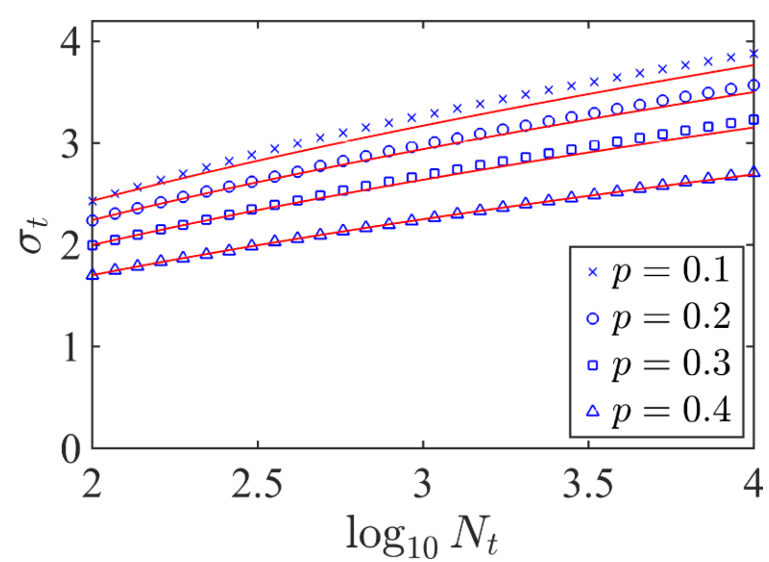
…. and distances are longer than a corresponding configuration model.

Node-Duplication Model

$$\langle L \rangle_t = 2(1-\eta)\ln\left(\frac{t+s+1}{s+1}\right) + C_1$$

$$\sigma_t^2 = 2(1-\eta)\ln\left(\frac{t+s+1}{s+1}\right) + C_2$$





Mean and variance scale logarithmically with the network size – i.e. it is a small world rather than ultra small world, like all other studied scale free networks. Also, unlike the configuration model the variance is also growing.

We studied the ***directed*** version too (relevant to citation networks).

Summary

We studied $P(L)$ the distribution of shortest path lengths (DSPL) in random networks using various techniques.

The challenge is twofold:

1. Demonstrating the utility of $P(L)$ in dynamical processes, and in describing the radial growth of shells around a random node. This part is independent of how $P(L)$ is calculated or measured.
2. Analytical results for $P(L)$ using the cavity method (equilibrium graphs), a Master equation (non-equilibrium graphs) and an exact approach to treat the subpercolating regime.

Our main claim is that the DSPL is a fundamental quantity although it did not receive enough attention.

Challenges: obtain the DSPL for other network ensembles.

Summary

Challenges: obtain the DSPL for other network ensembles.

References

1. E. Katzav, M. Nitzan, D. ben-Avraham, P.L. Krapivsky, R. Kühn, N. Ross and O. Biham, Analytical results for the DSPL in random networks, *EPL* **111**, 26006 (2015).

2. M. Nitzan, E. Katzav, R. Kühn and O. Biham, Distance distribution in configuration model networks, *Phys. Rev. E* **93**, 062309 (2016).

3. C. Steinbock, O. Biham and E. Katzav, The distribution of shortest path lengths in a class of node duplication network models, *Phys. Rev. E* **96**, 032301 (2017).

4. H. Bonneau, A. Hassid, O. Biham, R. Kühn and E. Katzav, Distribution of shortest cycle lengths in random networks, *Phys. Rev. E* **96**, 062307 (2017).

5. I. Tishby, O. Biham, E. Katzav and R. Kühn, Revealing the Micro-Structure of the Giant Clusters in Random Graph Ensembles, *Phys. Rev. E* **97**, 042318 (2018).

6. E. Katzav, O. Biham and A. Hartmann, The distribution of shortest path lengths in subcritical Erdős-Rényi networks, *Phys. Rev. E* **98**, 012301 (2018).

7. C. Steinbock, O. Biham and E. Katzav, Exact results for directed random networks that grow by node duplication, *Eur. Phys. J. B* **92**, 130 (2019).

8. I. Tishby, O. Biham, R. Kühn and E. Katzav, The mean and variance of the distribution of shortest path lengths of random regular graphs, *J. Phys. A* **55**, 265005 (2022).

9. B. Budnick, O. Biham and E. Katzav, The distribution of shortest path lengths on trees of a given size in subcritical Erdős-Rényi networks, submitted to Phys. Rev. E (2023)

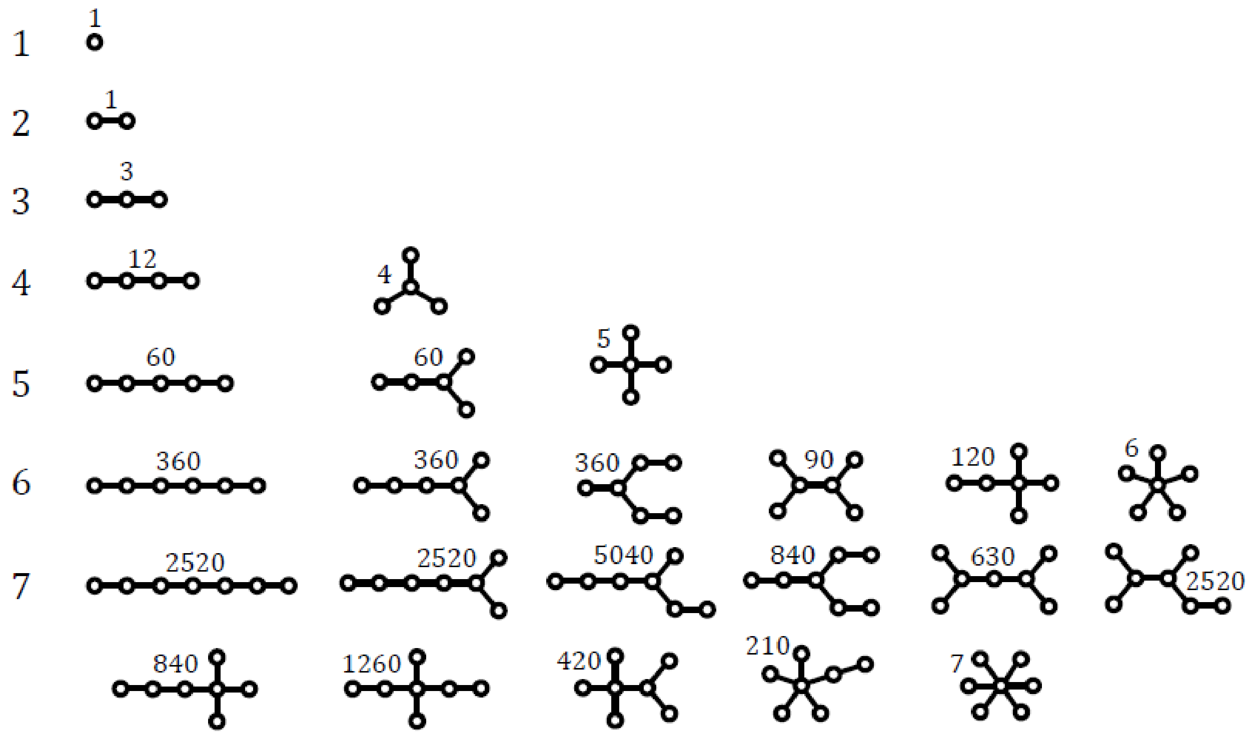
Subcritical case – $\mathrm{ER(N,p)},\ c < 1$

The probability that two nodes are connected is fairly low, namely

$$P(L < \infty) = \frac{c}{(1-c)N}$$

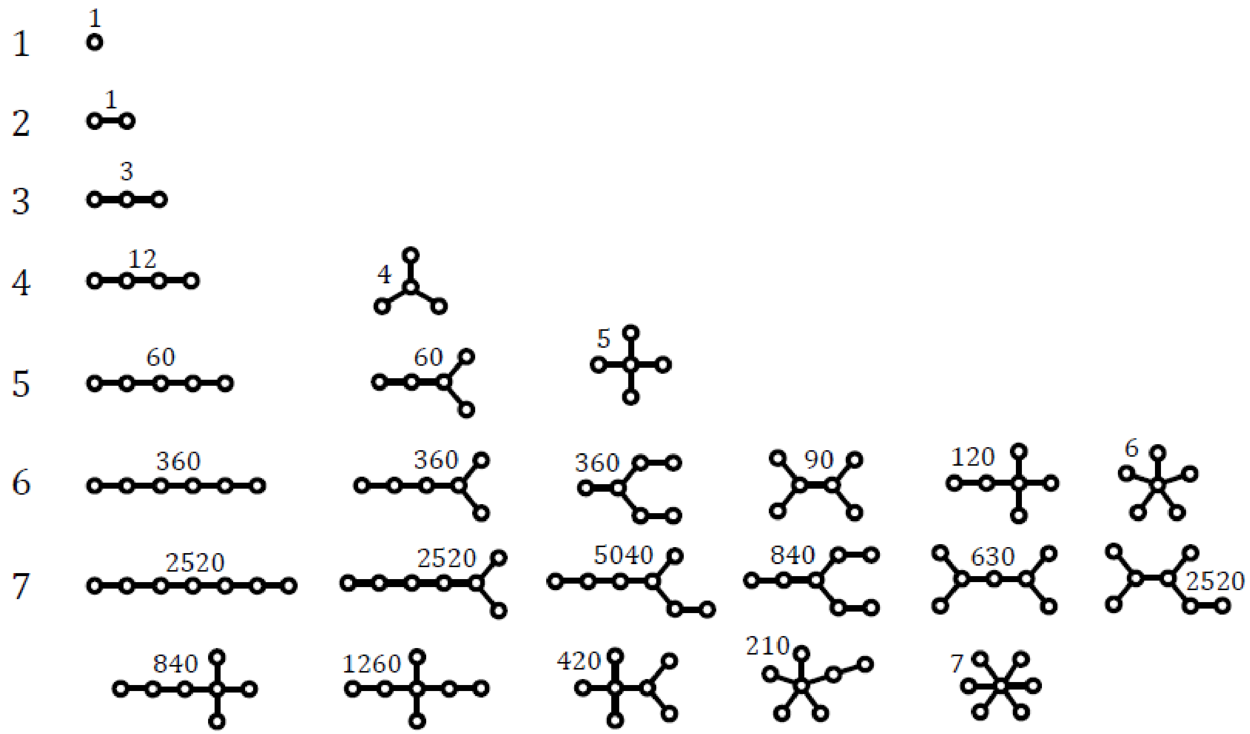
Thus, we focus on pairs for which $L < \infty$.

The basic methodology is a “topological expansion”



Topological expansion

For each basic topology we need to be able to calculate the DSPL, which can be quite challenging beyond the very basic ones. We have developed a powerful machinery to get those, classified by the size of the tree and the number of hubs $N\left(L=\ell|\tau=(h,A,\vec{b},s\right)$ (the number of topologies grow fast)



Topological expansion

For each basic topology we need to be able to calculate the DSPL, which can be quite challenging beyond the very basic ones. We have developed a powerful machinery to get those, classified by the size of the tree and the number of hubs $N\left(L = \ell | \tau = \left(h, A, \vec{b}, s\right)\right)$ - summarized below up to $s = 10$

We also use information about the relative weights of the different topologies

$P(L = \ell \mid L < \infty, S = s)$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$	$s = 9$	$s = 10$
$P_{FC}(L = 1 \mid L < \infty, S = s) =$	1	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{3}$	$\frac{2}{7}$	$\frac{1}{4}$	$\frac{2}{9}$	$\frac{1}{5}$
$P_{FC}(L = 2 \mid L < \infty, S = s) =$		$\frac{1}{3}$	$\frac{3}{8}$	$\frac{9}{25}$	$\frac{1}{3}$	$\frac{15}{49}$	$\frac{9}{32}$	$\frac{7}{27}$	$\frac{6}{25}$
$P_{FC}(L = 3 \mid L < \infty, S = s) =$			$\frac{1}{8}$	$\frac{24}{125}$	$\frac{2}{9}$	$\frac{80}{343}$	$\frac{15}{64}$	$\frac{56}{243}$	$\frac{28}{125}$
$P_{FC}(L = 4 \mid L < \infty, S = s) =$				$\frac{6}{125}$	$\frac{5}{54}$	$\frac{300}{2401}$	$\frac{75}{512}$	$\frac{350}{2187}$	$\frac{21}{125}$
$P_{FC}(L = 5 \mid L < \infty, S = s) =$					$\frac{1}{54}$	$\frac{720}{16807}$	$\frac{135}{2048}$	$\frac{560}{6561}$	$\frac{63}{625}$
$P_{FC}(L = 6 \mid L < \infty, S = s) =$						$\frac{120}{16807}$	$\frac{315}{16384}$	$\frac{1960}{59049}$	$\frac{147}{3125}$
$P_{FC}(L = 7 \mid L < \infty, S = s) =$							$\frac{45}{16384}$	$\frac{4480}{531441}$	$\frac{252}{15625}$
$P_{FC}(L = 8 \mid L < \infty, S = s) =$								$\frac{560}{531441}$	$\frac{567}{156250}$
$P_{FC}(L = 9 \mid L < \infty, S = s) =$									$\frac{63}{156250}$

Topological expansion

But this is just some preliminary work, because what really need to know is the composition of trees of different sizes in a given $\mathrm{ER}(N,p)$ network.

The number of trees of size s in such a network of size N is given by

$$T_s^N = N\binom{N}{s}s^{s-2}\left(\frac{c}{N}\right)^{s-1}\left(1-\frac{c}{N}\right)^{\binom{s}{2}-(s-1)}\left(1-\frac{c}{N}\right)^{s(N-s)}$$

or

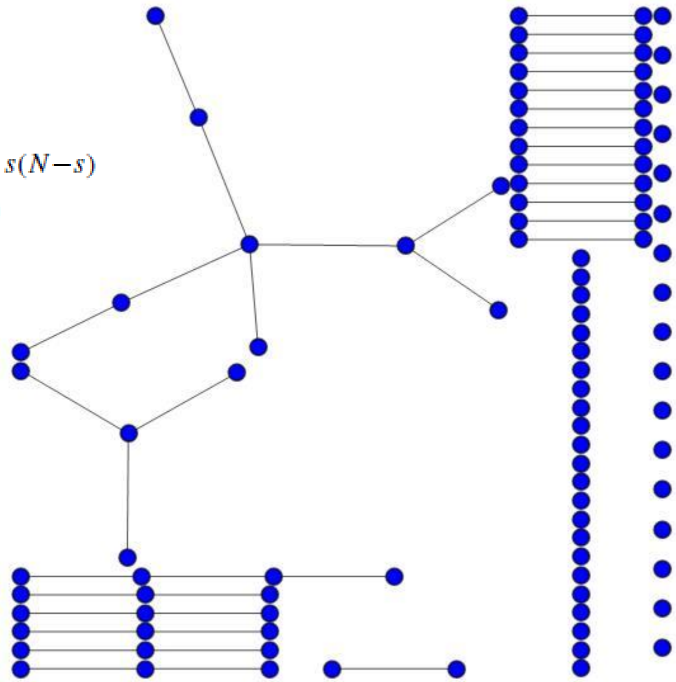
$$T_s^N \simeq N\frac{s^{s-2}c^{s-1}e^{-cs}}{s!}$$

Leading to

$$N_T(c) = \left(1-\frac{c}{2}\right)N$$

$$P_{\mathrm{FC}}(S=s) = \frac{2s^{s-2}c^{s-1}e^{-cs}}{(2-c)s!}$$

$$\langle S\rangle_{\mathrm{FC}} = \frac{2}{2-c}$$



N = 100, c = 0.7

Topological expansion

Now that we know $P_{\mathrm{FC}}(S=s) = \dfrac{2s^{s-2}c^{s-1}e^{-cs}}{(2-c)s!}$ and $\langle S \rangle_{\mathrm{FC}} = \dfrac{2}{2-c}$

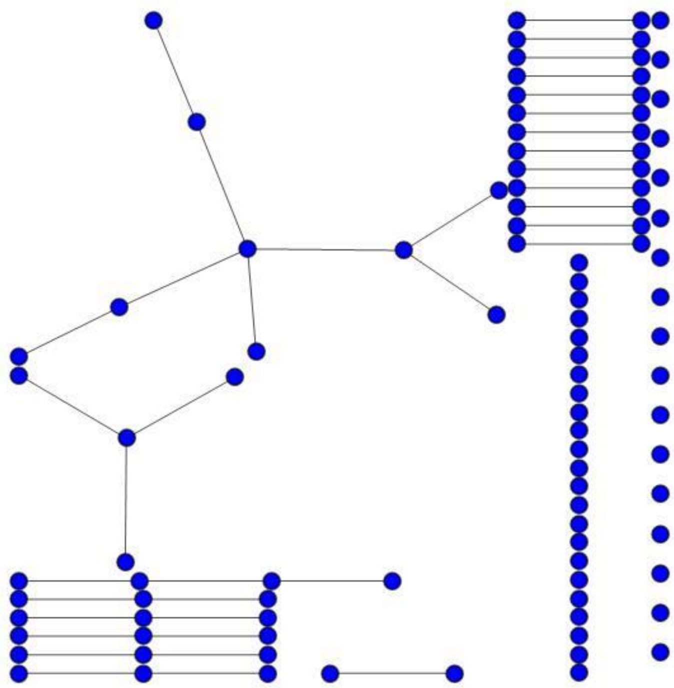
which corresponds to the probability that a randomly chosen tree is of size s.

But we also need the probability that a randomly chosen node belongs to a tree of size s

$$\widetilde{P}_{\mathrm{FC}}(S=s) = \frac{s}{\langle S \rangle_{\mathrm{FC}}} P_{\mathrm{FC}}(S=s)$$

$$\langle \widetilde{S} \rangle_{\mathrm{FC}} = \frac{\langle S^2 \rangle_{\mathrm{FC}}}{\langle S \rangle_{\mathrm{FC}}} = \frac{1}{1-c}$$

and in particular the probability that a randomly chosen edge belongs to a tree of size s



$$\widehat{P}_{\mathrm{FC}}(S=s) = \frac{\binom{s}{2} P_{\mathrm{FC}}(S=s)}{\left\langle \binom{S}{2} \right\rangle_{\mathrm{FC}}}$$ where $$\left\langle \binom{S}{2} \right\rangle_{\mathrm{FC}} = \frac{c}{(1-c)(2-c)}$$

$N = 100,\ c = 0.7$

Topological expansion

And we could combine everything together using

$$P(L=\ell|L<\infty)=\sum_{s=2}^{\infty}P(L=\ell|S=s)\widehat{P}(S=s)$$

Considering partial sums that take into account only trees up to size s, i.e. the quantity $P(L=\ell|L<\infty, S\leq s)$, and expanding in powers of c we obtained

$$P(L=2|S\leq 3)=(1-c)c\left(1-\frac{c}{3}+L\right)$$

$$P(L=2|S\leq 4)=(1-c)c\left(1-\frac{5c^2}{24}+L\right)$$

$$P(L=2|S\leq 5)=(1-c)c\left(1-\frac{9c^3}{25}+L\right)$$

$$P(L=2|S\leq 6)=(1-c)c\left(1-\frac{6517c^4}{14040}+L\right)$$

$$P(L=2|S\leq 7)=(1-c)c\left(1-\frac{30208c^5}{47943}+L\right)$$

L

$$P(L=2|S\leq 11)=(1-c)c\left(1-\frac{38523299328c^9}{12855460475}+L\right)$$

$$P(L=3|S\leq 4)=(1-c)c^2\left(1+\frac{c}{6}+L\right)$$

$$P(L=3|S\leq 5)=(1-c)c^2\left(1-\frac{3c^2}{25}+L\right)$$

$$P(L=3|S\leq 6)=(1-c)c^2\left(1-\frac{245c^3}{1404}+L\right)$$

$$P(L=3|S\leq 7)=(1-c)c^2\left(1-\frac{7936c^4}{15981}+L\right)$$

L

$$P(L=3|S\leq 11)=(1-c)c^2\left(1-\frac{329616179712c^8}{89988223325}+L\right)$$

The DSPL conditioned on tree size

We can actually play a trick and inverse this relation

$$P(L=\ell|L<\infty)=\sum_{s=2}^{\infty}P(L=\ell|S=s)\widehat{P}(S=s)$$

We know $\hat{P}(S=s)$ as well as $P(L=\ell|L<\infty)$, and we can thus try to obtain $P(L=\ell|S=s)$. Surprisingly, this works out nicely and we get

$$P(L=\ell|S=s)=\frac{\ell+1}{s^{\ell}}\frac{(s-1)!}{(s-\ell-1)!}$$

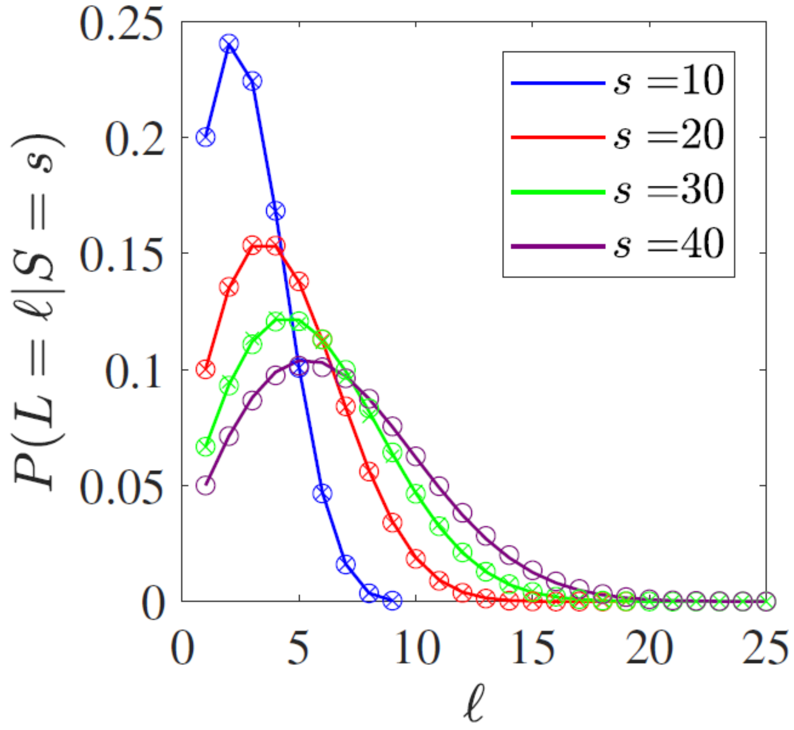
This expression does not depend on c!

Namely, once conditioning on the tree sizes all information about the mean degree c is lost. This comes from the fact that the ER construction actually samples trees of a given size s uniformly as labelled trees.

B. Budnick, O. Biham and E. Katzav, The distribution of shortest path lengths on trees of a given size in subcritical Erdős-Rényi networks, submitted to Phys. Rev. E (2023)

The DSPL conditioned on tree size

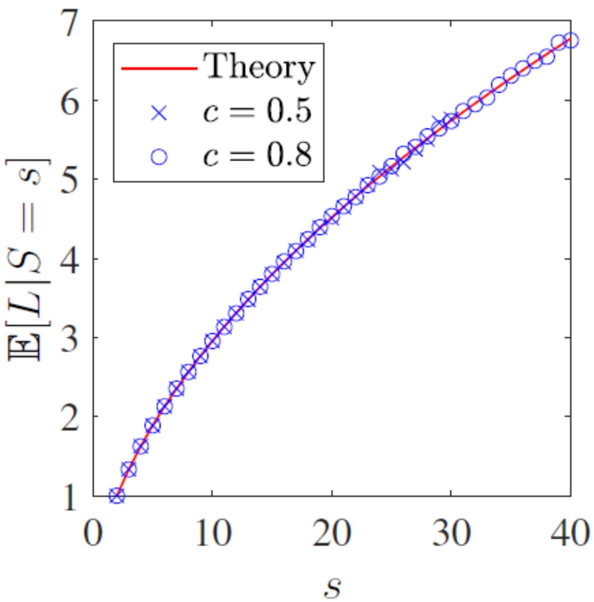
$$P(L = \ell | S = s) = \frac{\ell + 1}{s^{\ell}} \frac{(s-1)!}{(s-\ell-1)!}$$

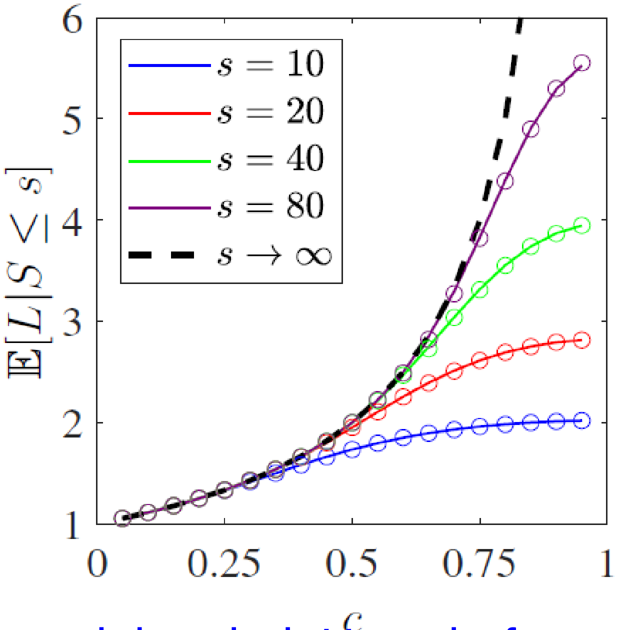


The DSPL conditioned on tree size

$$P(L = \ell | S = s) = \frac{\ell + 1}{s^\ell} \frac{(s-1)!}{(s - \ell - 1)!}$$

$$E[L|S = s] = \frac{s}{s-1}[e^s s^{-s} \Gamma(s+1, s) - 2] \simeq \sqrt{\frac{\pi}{2}}\sqrt{s} - \frac{4}{3} + O\left(\frac{1}{\sqrt{s}}\right)$$





This implies that typical trees grow isotropically, and that their Hausdorf dimension is ½ .

BA – numerical results

We still have no theory to predict the DSPL of the BA ensemble. However, we know that the degree distribution is a power-law with minor degree-degree correlations that vanish in the large N limit