Metric properties of subcritical Erdős-Rényi networks

Eytan Katzav,[1] Ofer Biham,[1] and Alexander K. Hartmann[2]

[1] *Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel*

[2] *Institute of Physics, Carl von Ossietzky University, 26111 Oldenburg, Germany*

Abstract

Networks which are fragmented into small disconnected components are prevalent in a large variety of systems. These include the secure communication networks of commercial enterprizes, goverment agencies and illicit organizations, as well as networks which suffered multiple failures, attacks or epidemics. The structural and statistical properties of such networks resemble those of random networks in the subcritical regime, below the percolation transition, which consist of finite clusters, whose sizes are non-extensive. Surprisingly, such networks do not exhibit the small-world property which is typical in supercritical random networks, where the mean distance between pairs of nodes scales logarithmically with the network size. Unlike supercritical networks whose structure has been studied extensively, subcritical networks have attracted relatively little attention. A special feature of these networks is that the statistical and geometric properties vary between different clusters and depend on their sizes and topologies. The overall statistics of the network can be obtained by a summation over all the clusters with suitable weights. We use a topological expansion to perform a systematic analysis of the distribution of shortest path lengths (DSPL) on clusters of given sizes and topologies in subcritical Erdős-Rényi (ER) networks. From this expansion we obtain an exact analytical expression for the DSPL of the entire subcritical network, in the asymptotic limit. The DSPL, which accounts for all pairs of nodes which reside on the same finite cluster (FC), is found to be a geometric distribution of the form $P_{\rm FC}(L = \ell | L < \infty) = (1-c)c^{\ell-1}$, where $c < 1$ is the mean degree. Using computer simulations we calculate the DSPL in subcritical ER networks of increasing sizes and confirm the convergence to this asymptotic result. We also obtain exact asymptotic results for the mean distance, $\langle L \rangle_{\rm FC}$, and for the standard deviation of the DSPL, $\sigma_{L,{\rm FC}}$, and show that the simulation results converge to these asymptotic results. Using the duality relations between subcritical and supercritical ER networks, we obtain the DSPL on the non-giant clusters of ER networks above the percolation transition.

I. INTRODUCTION

Network models provide a useful framework for the analysis of a large variety of systems which consist of interacting objects [1–3]. In these models, the objects are represented by nodes and the interactions between them are described by edges. A pair of nodes, i and j, may be connected via many different paths. The shortest among these paths are of particular importance because they provide the fastest and **often the** strongest interaction. Broadly speaking, one can distinguish between two major classes of networks: supercritical networks, which are tightly connected, and subcritical networks, which are loosely connected. Supercritical networks form a giant cluster which encompasses a macroscopic fraction of all the nodes, while the typical distances between pairs of nodes on the giant cluster scale logarithmically with the network size. Examples of such networks are the world-wide-web, social networks, and infrastructure networks of transportation, telephone, internet and electricity. In contrast, subcritical networks are fragmented into small clusters which do not scale with the overall network size. Examples of fragmented networks include secure networks with controlled access, such as the communication networks of commercial enterprizes, government agencies and illicit organizations [4]. Other examples include networks which suffered multiple failures, large scale attacks or epidemics, in which the remaining functional or uninfected nodes form small, isolated clusters [5, 6]. In spite of their prevalence, fragmented networks are of low visibility and have not attracted nearly as much attention as supercritical networks.

Random networks of the Erdős-Rényi (ER) type [7] are the simplest class of random networks and are used as a benchmark for the study of structure and dynamics in complex networks [8]. In an ER network of N nodes, each pair of nodes is independently connected with probability p, such that the mean degree is $c = (N-1)p$. The ensemble of such networks is denoted by $\mathrm{ER}[N,p]$. The degree distribution of these networks follows a Poisson distribution of the form

$$\pi(K=k) = \frac{e^{-c}c^k}{k!}. \tag{1}$$

ER networks exhibit a percolation transition at $c = 1$ above which there is a giant component, while below the transition the network consists of small, isolated components [8, 9]. To characterize the paths connecting random pairs of nodes, measures such as the diameter and the mean distance were studied [10–16]. For supercritical ER networks, it was shown that the mean distance, $\langle L \rangle$, scales like $\langle L \rangle \sim \ln N / \ln c$, in agreement with rigorous results, showing that percolating random networks are small world networks [10, 12]. For subcritical ER networks it was recently shown that the distribution of diameters over an ensemble of networks follows a Gumbel distribution of extreme values [15, 16]. This is due to the fact that in subcritical networks the diameter is obtained by maximizing the distances over all the small components. For supercritical networks, the entire distribution of shortest path lengths (DSPL) was calculated using various approximation techniques [5, 6, 17–23]. However, the DSPL of subcritical networks has not been studied.

The DSPL provides a natural platform for the study of dynamical processes on networks, such as diffusive processes, epidemic spreading, critical phenomena, synchronization, information propagation and communication [1–3, 25]. Thermal and dynamical processes on networks resemble those of systems with long range interactions in the sense that extensivity is broken and standard statistical physics techniques do not apply. Therefore, it is crucial to develop approaches for the study of dynamical processes on networks, that take into account their topological and geometrical properties, which are different from those of the Euclidean space. In fact, the DSPL provides exact solutions for various dynamical problems on networks. In the context of traffic flow on networks, the DSPL provides the distribution of transit times between all pairs of nodes, in the limit of low traffic load [26]. In the context of search processes, the DSPL determines the order in which nodes are explored in the breadth-first search protocol [26]. In the context of epidemic spreading, the DSPL captures the temporal evolution of the susciptible-infected (SI) epidemic, in the limit of high infection rate [25]. In the context of network attacks, the DSPL describes a generic class of violent local attacks, which spreads throughout the network [27].

The DSPL provides a useful characterization of empirical networks. For example, the DSPL of the protein network in Drosophila melanogaster was compared to the DSPL of a corresponding randomized network [28]. It was shown that proteins in this network are significantly farther away from each other than in the randomized network, providing useful biological insight. In the context of brain research, it was found that the DSPL and the distribution of shortest cycle lengths [29] determine the periods of oscillations in the activity of neural circuits [30, 31]. In essence, shortest paths and shortest cycles control the most important feedback mechanisms in these circuits, setting the characteristic time scales at which oscillations are sustained.

As mentioned above, in the asymptotic limit, $N \to \infty$, ER networks exhibit a percolation transition at $c = 1$. For $c < 1$, an ER network consists of finite clusters (FC) which are non-extensive with the network size, while for $c > 1$ a giant cluster (GC) is formed, which includes a finite fraction of the nodes in the network [8]. When a pair of nodes, i and j, resides on the same cluster, the distance, ℓ_{ij}, between them is defined as the length of the shortest path which connects them. **In case the edges carry no distance labels, as it is considered here, the length of a path is just the number of edges along a path.** When i and j reside on different clusters, there is no path connecting them and we define the distance between them to be $\ell_{ij} \equiv \infty$. We denote the probability distribution $P_{\rm FC}(L = \ell)$ as the DSPL over all $\binom{N}{2}$ pairs of nodes in the subcritical ER network. The probability that two randomly selected nodes reside on the same cluster and thus are at a finite distance from each other is denoted by $P_{\rm FC}(L < \infty)$.

Here we focus on the conditional DSPL between pairs of nodes which reside on the same cluster, denoted by $P_{\rm FC}(L = \ell | L < \infty)$, where $\ell = 1, 2, \dots, N-1$. This conditional DSPL satisfies

$$P_{\rm FC}(L = \ell | L < \infty) = \frac{P_{\rm FC}(L = \ell)}{P_{\rm FC}(L < \infty)}. \qquad (2)$$

In this paper we present a systematic analysis of the DSPL of subcritical ER networks using a topological expansion. We find that in the asymptotic limit the DSPL is given by a geometric distribution of the form $P_{\rm FC}(L = \ell | L < \infty) = (1-c)c^{\ell-1}$, where $c < 1$. Using computer simulations [32] we calculate the DSPL in subcritical ER networks of increasing sizes and confirm the convergence to this asymptotic result. We also show that the mean distance between pairs of nodes which reside on the same cluster is given by $\langle L \rangle_{\rm FC} = 1/(1-c)$. It is found that the mean distance scales linearly with the mean cluster size on which a random node resides. This is in contrast to supercritical random networks, which are small-world networks in which the mean distance scales logarithmically with N [10–12] or even ultrasmall world networks in which it scales sub-logarithmically with N [33]. Using duality relations connecting the non-giant components of supercritical ER networks to the corresponding subcritical ER networks [8, 34, 35], we obtain the DSPL of the non-giant components of the ER network above the percolation transition.

The paper is organized as follows. In Sec. II we review relevant properties of supercritical

ER networks and the percolation transition. In Sec. III describe recent results for the DSPL of supercritical networks. In Sec. IV we review some statistical properties of finite trees in subcritical ER networks, which are used in the analysis below. In Sec. V we present the topological expansion. In Sec. VI we use the topological expansion to calculate the degree distribution on ensembles of trees of given sizes and topologies. In Sec. VII we use the topological expansion to calculate the DSPL on emsembles of trees of given sizes and topologies and to obtain an exact analytical expression for the DSPL in the asymptotic limit. In Sec. VIII we calculate the mean and standard deviation of the DSPL. The results are summarized and discussed in Sec. IX. In Appendix A we present the derivation of a formula for $P_{\rm FC}(L < \infty)$, namely the probability that two random nodes which reside on the finite clusters will reside on the same cluster.

II. SUPERCRITICAL NETWORKS: THE GIANT CLUSTER

The ER network emsemble is a special case of a broader class of network models called configuration model networks. The configuration model ensemble is a maximum entropy ensemble of networks, under the condition that the degree distribution, $P(K = k)$ is imposed [2, 13]. To construct such a network of N nodes, one draws the degrees of all the nodes from $P(K = k)$, producing a degree sequence of the form $\{k_i\}_{i=1,\dots,N}$ (where $\sum k_i$ must be even). The nodes are then connected to each other randomly according to the degree sequence such that no degree-degree correlations are formed between pairs of adjacent nodes. **When generating such graphs in the computer, self-edges or multiple edges might be created. Such realizions have to be discarded, not "repaired", in order to generate the graphs with the correct weights [36].** The ER network ensemble, which exhibits a Poisson degree distribution, is a maximum entropy ensemble, under the condition that the mean degree $\langle K \rangle = c$ is fixed.

Consider a configuration model network of N nodes with a given degree distribution $P(K = k)$. When nodes are sampled randomly from the network the distribution of their degrees is $P(K = k)$. However, nodes which are sampled as random neighbors of random nodes follow a modified degree distribution, which takes the form

$$\widetilde{P}(K = k) = \frac{k}{\langle K \rangle} P(K = k). \tag{3}$$

This is due to the fact that such nodes are selected proportionally to their degrees. For ER networks the Poisson distribution, given by Eq. (1), satisfies

$$\tilde{\pi}(K = k) = \frac{k}{c}\pi(K = k) = \pi(K = k - 1). \tag{4}$$

This means that the probability that a random neighbor of a random node is of degree k is equal to the probability that a random node is of degree $k - 1$. Moreover, it implies that in ER networks the degree distribution of nodes selected as a random neighbors of a random reference node, in the reduced network from which the reference node is removed, is identical to the degree distribution of a random node in the complete network. The generating function of the Poisson distribution is

$$G_0(x) = \sum_{k=0}^{\infty} \pi(K = k)x^k, \tag{5}$$

The probability that a random node resides on the giant cluster is denoted by g. In case that a giant cluster exists, $g > 0$, while in case that there is no giant cluster, $g = 0$. In the thermodynamic limit, $N \to \infty$, the probability g is given as a solution of the self-consistent equation [1]

$$1 - g = G_0(1 - g). \tag{6}$$

The left hand side of this equation is the probability that a random node, i, does not reside on the giant cluster. The right hand side represents the probability that none of its neighbors of such node resides on the giant cluster in the reduced network, which does not include the node i. Clearly, $g = 0$ is always a solution of Eq. (6). An ER network exhibits a giant cluster if Eq. (6) also has a non-trivial solution. Inserting the function $G_0(x)$ from Eq. (5) into Eq. (6) and performing the summation, one obtains $1 - g = \exp(-cg)$ [8]. Solving this equation, one finds that $g = 0$ for $c \leq 1$ and $g = 1$ for $c > \ln N$. For intermediate values of c, in the range of $1 < c < \ln N$, the fraction of nodes $g = g(c)$ which reside on the giant cluster is given by

$$g(c) = 1 + \frac{\mathcal{W}(-ce^{-c})}{c}, \tag{7}$$

where $\mathcal{W}(x)$ is the Lambert W function [37]. This means that ER networks exhibit a percolation transition at $c = 1$, such that for $c < 1$ the network consists only of finite components while for $c > 1$ there is a giant cluster. At a higher value of the connectivity, namely at $c = \ln N$, there is a second transition, above which the giant cluster encompasses the entire network and there are no non-giant components.

III. THE DSPL OF SUPERCRITICAL ERDŐS-RÉNYI NETWORKS

Consider a pair of random nodes, i and j, in a supercritical ER network of N nodes. The probability that both of them reside on the giant cluster is g^2, the probability that one of them resides on the giant cluster and the other resides on one of the finite clusters is $2g(1-g)$, while the probability that both reside on finite clusters is $(1-g)^2$. Assuming both nodes reside on the giant cluster, they may be connected to each other by a large number of paths. In case that one of them resides on the giant cluster and the other resides on one of the finite components, the distance is $\ell_{ij} = \infty$. In case that both of them reside on finite clusters, a path between them exists only in the less probable case that they reside on the same finite cluster. The finite clusters are trees and therefore the shortest path between any pair of nodes is unique. Therefore, the DSPL of a supercritical ER network can be expressed by

$$\begin{aligned} P(L > \ell) = g^2 P_{\rm GC}(L > \ell) + (1-g)^2 P_{\rm FC}(L < \infty) P_{\rm FC}(L > \ell | L < \infty) \\ + 2g(1-g) + (1-g)^2 \left[1 - P_{\rm FC}(L < \infty) \right], \end{aligned} \tag{8}$$

where the first term accounts for the DSPL on the giant cluster, the second term account for the DSPL of the finite clusters and the last two term account for the probability that i and j reside on different clusters. For $1 < c < \ln N$, where $0 < g < 1$, the network consists of a combination of a giant and finite clusters, where the weight of the finite cluster increases as c is decreased. In this regime of coexistence, the calculation of the DSPL on the giant cluster is a non-trivial task. This is due to the fact that in this regime the giant cluster is a more complicated geometrical object. Its degree distribution deviates from the Poisson distribution and it exhibits degree-degree correlations. This DSPL was recently analyzed using a theoretical framework which is based on the shell structure around a random node [38].

For $c > \ln N$ the giant cluster encompasses the entire network and $g = 1$. In this regime only the first term is required and the network can be considered as a single component. Under these conditions the DSPL can be expressed as a product of the form

$$P(L > \ell) = \prod_{\ell'=1}^{\ell} P(L > \ell' | L > \ell' - 1), \tag{9}$$

where $P(L > \ell | L > \ell - 1)$ is the conditional probability that the distance between a random pair of nodes is larger than ℓ conditioned on it being larger than $\ell - 1$

A path of length ℓ from node i to node j can be decomposed into a single edge connecting node i and node $r \in \partial_i$ (where ∂_i is the set of all nodes directly connected to i), and a shorter path of length $\ell - 1$ connecting r and j. Thus, the existence of a path of length ℓ between nodes i and j can be ruled out if there is no path of length $\ell - 1$ between any of the nodes $r \in \partial_i$, and j. For sufficiently large networks, the argument presented above translates into the recursion equation [22]

$$P(L > \ell | L > \ell - 1) = G_0[P(L > \ell - 1 | L > \ell - 2)], \tag{10}$$

where the generating function $G_0(x)$ is given by Eq. (5). The case of $\ell = 1$ deserves special attention. On a network of size N (sufficiently large), the probability that two random nodes are not connected is given by [22]

$$P(L > 1 | L > 0) \simeq 1 - \frac{c}{N-1} + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{11}$$

The recursion equations provide a good approximation for the DSPL of supercritical ER networks [21–23], for values of $c > 1$ which are not too close to the percolation threshold. However, no exact result for the DSPL of supercritical ER networks is known. Interestingly, for random regular graphs, in which all the nodes are of the same degree, $k = c \geq 3$, there is an exact result for the DSPL, which can be expressed by a Gompertz distribution [24] of the form [11, 22]

$$P(L > \ell) = \exp\left[-\beta(e^{b\ell} - 1)\right], \tag{12}$$

where $\beta = c/[N(c-2)]$ and $b = \ln(c-1)$. For a supercritical ER network with mean degree c which is sufficiently far above the percolation threshold, the DSPL is qualitatively similar to the DSPL of a random regular graph with degree $\lfloor c + 1/2 \rfloor$. Here, $\lfloor x \rfloor$ is the integer part of x and thus $\lfloor x + 1/2 \rfloor$ is the nearest integer to x. Unlike random regular graphs in which all the nodes are of the same degree, in supercritical ER networks, which follow the Poisson degree distribution, the shortest path length between a pair of nodes is correlated with the degrees of these nodes. The correlation is negative, namely nodes of high degrees tend to be closer to each other than nodes of low degrees. Another simplifying feature of random regular graphs with $c \geq 3$ is that the giant cluster encompasses the entire network ($g = 1$) and thus all pairs of nodes are connected by finite paths. Since ER networks with $1 < c < \ln N$ consist of a combination of a giant cluster and finite clusters, the DSPL exhibits a non-zero asymptotic tail and its calculation is more difficult.

The DSPL on the finite clusters in supercritical ER networks is a sub-leading term in the overall DSPL, which involves a fraction of $(1-g)^2 P_{\rm FC}(L < \infty)$ from the $\binom{N}{2}$ pairs of nodes in the network. The factor of $(1-g)^2$ accounts for the fraction of pairs which reside on finite clusters, while the fraction of those pairs which reside on the same cluster is given by $P_{\rm FC}(L < \infty)$. Except for the vicinity of the percolation transition, which occurs at $c = 1$, this amounts to a small fraction of all pairs of nodes.

In the asymptotic limit, ER networks exhibit duality with respect to the percolation threshold. In a supercritical ER network of N nodes the fraction of nodes which belong to the giant cluster is $g(c)$ [Eq. (7)], while the fraction of nodes which belong to the finite clusters is $f(c) = 1 - g(c)$. Thus, the network which consists of the finite clusters is of size $N' = Nf(c)$. It can be shown **COMMENT:** *citation ?* that this network is an ER network whose mean degree is $c' = cf(c)$, where $c' < 1$. It means that the DSPL and other properties of the finite clusters of supercritical ER networks can be obtained from the analysis of the dual subcritical networks.

IV. SUBCRITICAL NETWORKS: PROPERTIES OF THE FINITE TREES

In the analysis presented below we use the fact that clusters forming in an ER network for $c < 1$ are almost surely trees, namely the expected number of cycles is non-extensive [8]. The expected number of tree components of size s in a network of size N is denoted by T_s^N, so that $\sum_{s \geq 1} s T_s^N = N$. In these trees we define all the nodes of degree $k \geq 3$ as *hubs* and all the nodes of degree $k = 1$ as *leaves*. Linear chains of nodes which have a hub on one side and a leaf on the other side are referred to as *branches*, while chains which have hubs on both sides are referred to as *arms*. In Fig. 1 we illustrate the structure of an ER network of size $N = 100$ and $c = 0.9$. It consists of 33 isolated nodes, 9 dimers, two chains of three nodes, two chains of four nodes, one tree with a single hub and four branches, one tree with two hubs and two larger trees of 10 and 14 nodes.

Using the theory of branching processes, it was shown that [8, 9]

$$T_s^N = N \binom{N}{s} s^{s-2} \left(\frac{c}{N}\right)^{s-1} \left(1 - \frac{c}{N}\right)^{\binom{s}{2}-(s-1)} \left(1 - \frac{c}{N}\right)^{s(N-s)}, \tag{13}$$

where the binomial coefficient accounts for the number of ways to pick s nodes out of N in order to form a cluster of size s and the factor of s^{s-2} is the number of distinct tree structures which can be constructed from s distinguishable nodes [39]. The factor of $(c/N)^{s-1}$ accounts for the probability that the s nodes of the cluster will be connected by $s-1$ edges. The next term is the probability that there are no other edges connecting pairs of nodes in the cluster, while the last term is the probability that there are no edges connecting nodes in the clusters with nodes in the rest of the network. For $s \ll N$ one can approximate the binomial coefficient by $N^s/s!$ and obtain **COMMENT:** *Shouldn't it be N^2 because there is the N and the N^s/N^{s-1} above? Or is the factor N above wrong, it is not explained anyway?*

$$T_s^N = N \frac{s^{s-2} c^{s-1}}{s!} \left(1 - \frac{c}{N}\right)^{s(N-s)+\binom{s}{2}-(s-1)}. \tag{14}$$

Since we consider subcritical ER networks, for which $c < 1$, unless the network is extremely small the condition $c \ll N$ is satisfied. Therefore, one can approximate the last term in Eq. (14) by an **exponential**, and obtain

$$T_s^N = N \frac{s^{s-2} c^{s-1} e^{-cs}}{s!} \exp\left[\frac{c(s^2 + 3s - 2)}{2N}\right]. \tag{15}$$

Finally, in the asymptotic limit of $N \to \infty$, **the exponential converges towards one and** the expression for the expected number of tree components of s nodes is reduced to [8, 9]

$$T_s^N \simeq N \frac{s^{s-2} c^{s-1} e^{-cs}}{s!}. \tag{16}$$

In the limit of large s, one can use the Stirling approximation $s! = \sqrt{2\pi s}(s/e)^s$ and obtain

$$T_s^N \simeq \frac{N}{\sqrt{2\pi}c}\frac{e^{-s/s_{\rm max}}}{s^{5/2}}, \tag{17}$$

where the cutoff parameter $s_{\rm max}$ is given by

$$s_{\rm max} = \frac{1}{\ln\left(\frac{1}{ce^{1-c}}\right)}. \tag{18}$$

As the percolation threshold is approached from below, for $c \to 1^-$, the cutoff parameter diverges, according to $s_{\rm max} \sim 1/(1-c)^2$. The expected number of trees of size s per node, obtained from Eq. (17) scales like $T_s^N/N \propto s^{-\tau}$, where $\tau = 5/2$. This is in agreement with the critical cluster size distribution on regular lattices above the upper critical dimension of $D = 6$, where τ is the Fisher exponent [40], exemplifying the connection between percolation transitions on random networks and regular lattices of high dimensions.

Selecting two random nodes in the network, the probability that they reside on the same cluster is

$$P_{\rm FC}(L < \infty) = \frac{U(N,c)}{\binom{N}{2}}, \tag{19}$$

where $U(N,c)$ is the expected number of pairs of nodes in the network which reside on the same cluster. It is given by

$$U(N,c) = \sum_{s \ge 1} \binom{s}{2} T_s^N \tag{20}$$

In Appendix A we evaluate $U(N,c)$ and show that

$$P_{\rm FC}(L < \infty) = \frac{c}{(1-c)N}. \tag{21}$$

Using this result and the fact that the first two terms of $P_{\rm FC}(L = \ell)$ are known exactly, namely $P_{\rm FC}(L = 1) = p$ and $P_{\rm FC}(L = 2) = (1-p)[1-(1-p^2)^{N-2}]$ [21], we obtain that $P_{\rm FC}(L = 1|L < \infty) = 1 - c$ and $P_{\rm FC}(L = 2|L < \infty) = c(1-c)$.

The total number of tree components in a network of N nodes and $c < 1$ is denoted by

$$N_T(c) = \sum_{s=1}^{N} T_s^N. \tag{22}$$

Carrying out the summation we obtain **COMMENT:** *State which of the many approximations of T_s^N above was used here, also the special summation formula which was used* that for $0 \le c \le 1$

$$N_T(c) = \left(1 - \frac{c}{2}\right) N, \tag{23}$$

namely $N_T(c)$ is a linear, monotonically decreasing function of c, where $N_T(c=0) = N$ and $N_T(c=1) = N/2$. The mean tree size is thus given by

$$\langle S \rangle_{\rm FC} = N/N_T(c) = \frac{2}{2-c}, \tag{24}$$

which does not diverge as c approaches the percolation threshold. Using **Eqs. (16) and (23)**, we can write down the disribution of tree sizes, which takes the form

$$P_{\rm FC}(S=s) = T_s^N/N_T(c) = \frac{2s^{s-2}c^{s-1}e^{-cs}}{(2-c)s!}. \tag{25}$$

In various processes on networks clusters are selected by by drawing random nodes and choosing the clusters on which they reside. The probability that a randomly selected node resides on a tree of size s is given by

$$\widetilde{P}_{\rm FC}(S=s) = \frac{s}{\langle S \rangle_{\rm FC}} P_{\rm FC}(S=s). \tag{26}$$

The mean of this distribution is

$$\langle \tilde{S} \rangle_{\rm FC} = \frac{\langle S^2 \rangle_{\rm FC}}{\langle S \rangle_{\rm FC}} = \frac{1}{1-c}. \tag{27}$$

Thus, as $c \to 1^-$, the mean tree size on which a random node resides diverges.

Consider a random pair of nodes which reside on the same cluster. The probability that they reside on a cluster of size s is given by

$$\widehat{P}_{\rm FC}(S=s) = \frac{\binom{s}{2} P_{\rm FC}(S=s)}{\left\langle \binom{S}{2} \right\rangle_{\rm FC}}. \tag{28}$$

Evaluating the denominator we obtain

$$\left\langle \binom{S}{2} \right\rangle_{\rm FC} = \frac{c}{(1-c)(2-c)}. \tag{29}$$

The mean of $\widehat{P}_{\rm FC}(S=s)$ is found to be

$$\langle \widehat{S} \rangle_{\rm FC} = \frac{2-c}{(1-c)^2}, \tag{30}$$

which diverges quadratically as $c \to 1^-$.

In the next Section we introduce the topological expansion method. In this approach, for each cluster size, s, we identify all the possible tree topologies supported by s nodes, starting from the linear chain which does not include any hubs, followed by single-hub topologies, double-hub topologies and higher order topologies which include multiple-hubs. For each tree topology, we calculate its relative weight among all possible tree topologies of the same size. A special property of tree topologies is that each pair of nodes is connected by a single path. Therefore, in subcritical ER networks the shortest path between any pair of nodes is, in fact, the only path between them. Using this property we calculate the DSPL for each tree topology, and use the weights to obtain the DSPL over all the clusters which consist of up to s nodes.

V. THE TOPOLOGICAL EXPANSION

Consider a tree which includes h hubs. Embedded in this tree, there is a backbone tree, which consists only of the h hubs and the $h-1$ arms which connect them. The structure of the backbone tree is described by its adjacency matrix, A. This is a symmetric $h \times h$ matrix in which $A_{ij} = 1$ if hubs i and j are connected by an arm and 0 otherwise. The degrees of the hubs in the backbone tree are denoted by the vector

$$\vec{a} = (a_1, a_2, \dots, a_h), \tag{31}$$

where

$$a_i = \sum_{j=1}^{h} A_{ij}. \tag{32}$$

The structure of the branches is described by the vector

$$\vec{b} = (b_1, b_2, \dots, b_h), \tag{33}$$

where b_i is the number of branches connected to the $i^{\rm th}$ hub. The total number of branches in a tree is given by

$$b = \sum_{i=1}^{h} b_i. \tag{34}$$

The topology of a tree is fully described by the adjacency matrix, A, of its backbone tree and its branch vector $\vec{b}$. We denote such tree topology by

$$\tau = (h, A, \vec{b}). \tag{35}$$

In this classification, the linear chain structure is denoted by $\tau = (0, \cdot, 2)$. It has no nodes and thus $h = 0$. The matrix A is not defined and replaced by the "$\cdot$" sign. The linear chain has two leaf nodes and it is thus considered as a tree with two branches. A tree which includes a single hub with $b \geq 3$ branches is denoted by $\tau = (1, 0, b)$. Here, the matrix $A = 0$ is a scalar. A tree which includes two hubs with a branch vector $\vec{b} = (b_1, b_2)$, is denoted by $\tau = (2, A, \vec{b})$, where A is a 2×2 matrix of the form

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{36}$$

A tree which consists of a chain of three hubs with a branch vector $\vec{b} = (b_1, b_2, b_3)$, is denoted by $\tau = (3, A, \vec{b})$, where A is a 3×3 matrix of the form

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \tag{37}$$

A tree which includes four hubs, consisting of one central hub surrounded by three peripheral hubs is denoted by $\tau = (4, A, \vec{b})$ where

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{38}$$

and $\vec{b} = (b_1, b_2, b_3, b_4)$. In Fig. 2 we present all the possible topologies of the backbone tree that can be obtained with up to six hubs.

The topological expansion is performed such that the s^{th} order consists of all possible tree topologies which can be assembled from s nodes. Since each branch consists of at least one node, the number of branches in a tree which consists of s nodes and include h hubs must satisfy

$$b \leq s - h. \tag{39}$$

Unlike the branches, each arm may either consist of a single edge between the adjacent hubs or include one or more intermediate nodes. The number of arms connecting the h hubs in the tree is $h-1$. The degree of each hub is given by the sum of the number of branches and the number of arms connected to it. While each branch is connected to only one hub, each arm is connected to two hubs, one on each side. Recalling that the degree of each hub is $k \geq 3$ we find that $2(h-1)+b \geq 3h$. Thus, the number of branches in a tree which includes h hubs must satisfy

$$b \geq h + 2. \tag{40}$$

Combining Eqs. (39) and (40) we obtain a condition on the minimal tree size which may include h hubs. It takes the form

$$s \geq 2h + 2. \tag{41}$$

We thus obtain a classification of all tree structures which can be assembled from s nodes, where $s \geq 2$. For $s = 2, 3$ the linear chain is the only possible topology. Higher order topologies, which exist for $s \geq 4$, include at least one hub. In a tree of size $s \geq 4$, the number of hubs may take values in the range

$$h = 1, 2, \dots, \left\lfloor \frac{s}{2} - 1 \right\rfloor . \tag{42}$$

For each choice of h, the number of branches may take any value in the range

$$b = h + 2, h + 3, \dots, s - h. \tag{43}$$

In Fig. 3 we illustrate the possible values of b in a tree of h hubs, which consists of s nodes, given by Eq. (43), **i.e., the range is bounded from below by $b = h + 2$ and from above by $b = s - h$.** Combinations of (h, b) which are possible in a tree of $s = 12$ nodes are marked by full circles, while combination which exist only in larger trees are marked by empty circles. The possible topologies of backbone trees which consist of h hubs, for $h = 1, 2, \ldots, 5$, are shown at the bottom. For $h = 1$ the backbone tree consists of a single node. For $h = 2$ the backbone tree is a dimer. For $h = 3$ the backbone tree is a linear chain of three nodes. For $h = 4$ there are two possible tree topologies: a linear chain and a tree which consists of a central node surrounded by three peripheral nodes. The number of non-isomorphic tree topologies, $n(h)$, which can be assembled from h nodes quickly increases as a function of h. For example, the values of $n(h)$ for $h = 1, 2, \ldots, 13$ are 1, 1, 1, 2, 3, 6, 11, 23, 47, 106, 235, 551 and 1301, respectively [41]. An efficient algorithm, often referred to as the WRON algorithm, for generating all the tree topologies known as free trees, that can be assembled from h nodes, is presented in Refs. [42, 43]. A list of all possible tree topologies up to $h = 13$ is presented in Ref. [41]. Other web resources include enumeration of such tree topologies up to $h = 20$ [44].

Each one of the possible topologies of the backbone tree is represented by its adjacency matrix, A, which is an $h \times h$ matrix. The degrees of the hubs in the backbone tree are given by $\vec{a} = (a_1, a_2, \ldots, a_h)$. Since the degrees of the hubs, which are given by $k_i = a_i + b_i$, must satisfy the condition $k_i \geq 3$, the components of the branch vector, $\vec{b} = (b_1, b_2, \ldots, b_h)$ satisfy the condition

$$b_i \geq (3 - a_i)\theta(3 - a_i), \tag{44}$$

where $\theta(x)$ is the Heaviside function which satisfies $\theta(x) = 1$ for $x > 0$ and 0 for $x \leq 0$. The number of branches required to satisfy this condition is $h + 2$. In case that $b > h + 2$, the remaining $b - h - 2$ branches can be divided in many different ways between the h hubs. The number of possible partitions of x identical objects among y distinguishable boxes is given by the multiset coefficient [45]

$$\left(\!\!\binom{y}{x}\!\!\right) = \binom{x + y - 1}{x} = \binom{x + y - 1}{y - 1}. \tag{45}$$

Therefore, the number of ways different tree topologies which emanate from a single topology of a backbone tree of h hubs is

$$\left(\!\!\binom{h}{b - h - 2}\!\!\right) = \binom{b - 3}{h - 1}. \tag{46}$$

Consider all the tree topologies which can be assembled from s nodes. The weight of each tree topology, τ, is given by the number of ways to distribute s indistinguishable nodes among its branches and arms. We denote this weight by $W(\tau; s)$. In the case of a tree which consists of a linear chain of s nodes, there are no degrees of freedom, and therefore its weight is

$$W[\tau = (0, \cdot, 2); s] = 1. \tag{47}$$

The weight of a tree of s nodes which consists of a single hub and b branches is given by

$$W[\tau = (1, 0, b); s] = \binom{s-2}{b-1}. \tag{48}$$

Here the binomial factor counts the number of possibilities to distribute $s-1$ nodes between the b branches, such that each branch consists of at least one node. The weight of a tree with two hubs is

$$W[\tau = (2, A, \vec{b}); s] = \binom{s-2}{b}. \tag{49}$$

In this case the binomial coefficient accounts for the number of ways to distribute $s-2$ nodes between the b branches and one arm, where each branch includes at least one node.

In general, the weight of a tree structure consisting of s nodes, h hubs (which are connected by $h-1$ arms) and a branch vector $\vec{b}$ is

$$W(\tau = (h, A, \vec{b}); s) = \binom{s-2}{h+b-2}. \tag{50}$$

This result can be understood as follows. Among the s nodes, h nodes are fixed as hubs while each one of the b branches includes at least one node. Therefore, there are $x = s-h-b$ nodes which can be distributed among the $y = b+h-1$ branches and arms. Using Eq. (45) for the number of possible divisions of x objects among y boxes yields the result of Eq. (50).

The contribution of each tree topology to statistical properties of the network such as the degree distribution and the DSPL also depends on its symmetry. To account for the effect of the symmetry, we define the symmetry factor

$$X(\tau) = \frac{1}{|\mathrm{Aut}(\tau)|}, \tag{51}$$

where $\mathrm{Aut}(\tau)$ is the automorphism group of τ [8], namely all the transformations that leave τ unchanged. It can be expressed as a product of the form

$$|\mathrm{Aut}(\tau)| = |\mathrm{Aut}(A)||\mathrm{Aut}(\vec{b})|, \tag{52}$$

where Aut(A) is the automorphism group of the backbone tree, which consists of the hubs alone, and Aut($\vec{b}$) is the automorphism group of the branches. While $|\mathrm{Aut}(A)|$ depends on the overall symmetry of the backbone tree, $|\mathrm{Aut}(\vec{b})|$ is given by

$$|\mathrm{Aut}(\vec{b})| = \prod_{i=1}^{h} b_i!. \tag{53}$$

For example, in the case of a linear chain of s nodes,

$$X[\tau = (0, \cdot, 2)] = \frac{1}{2}. \tag{54}$$

For a tree consisting of a single hub of b branches

$$X[\tau = (1, 0, b)] = \frac{1}{b!}, \tag{55}$$

while for a tree which includes two hubs with b_1 and b_2 branches,

$$X[\tau = (2, A, \vec{b})] = \frac{1}{2b_1!b_2!}. \tag{56}$$

For a tree which consists of a central hub surrounded by three peripheral hubs

$$X[\tau = (4, A, \vec{b})] = \frac{1}{3b_1!b_2!b_3!b_4!}. \tag{57}$$

VI. THE DEGREE DISTRIBUTION

In this Section we show how to use the topological expansion in order to express the degree distribution $P(K = k)$ as a composition of the contributions of the different tree topologies. In this case the asymptotic form is known to be the Poisson distribution, $\pi(K = k)$, which enables us to validate the method.

Consider a tree which consists of s nodes, whose degree sequence is given by $k_i, i = 1, 2, \ldots, s$. Since a tree of s nodes includes $s - 1$ edges, the sum of these degrees satisfies

$$\sum_{i=1}^{s} k_i = 2(s-1). \tag{58}$$

We denote the number of nodes of degree k by $N(K=k)$, where

$$\sum_{k=1}^{s-1} N(K=k) = s, \tag{59}$$

reflecting the fact that in a tree of size $s > 1$ the degrees of all nodes satisfy $k \geq 1$. Using this notatin, Eq. (58) can be written in the form

$$\sum_{k=1}^{s-1} kN(K=k) = 2(s-1). \tag{60}$$

Combining the last two equations **to eliminate** s, we obtain

$$N(K=1) = 2 + \sum_{k=3}^{s-1} (k-2)N(K=k). \tag{61}$$

This result reflects the fact that any tree includes at least two leaf nodes and provides a relation between the degrees of the hubs and the number of leaf nodes in a tree. The number of nodes of degree $k=2$ can be obtained from **COMMENT:** *shouldn't this just be* $N(K=2) = s - N(K=1) - \sum_{k=3}^{s-1} N(K=k)$*?*

$$N(K=2) = s - N(K=1) - \sum_{k=3}^{s-1} (k-2)N(K=k). \tag{62}$$

The topology of each tree structure can be described by $\tau = (h, A, \vec{b})$, where

$$h = \sum_{k=3}^{s-1} N(K=k) \tag{63}$$

is the number of hubs. The degrees of the hubs are given by

$$\begin{aligned} k_1 &= a_1 + b_1 \\ k_2 &= a_2 + b_2 \\ &\vdots \\ k_h &= a_h + b_h, \end{aligned} \tag{64}$$

where a_i is the number of arms and b_i is the number of branches which are connected to hub i. The number of leaf nodes, with degree $k=1$ is given by $b = \sum_{i=1}^{h} b_i$. The remaining $s-h-b$ nodes are of degree $k=2$.

The number of nodes of degree k in a linear chain of s nodes is given by

$$N(K = k) = 2\delta_{k,1} + (s - 2)\delta_{k,2}, \tag{65}$$

where $\delta_{k,k'}$ is the Kronecker delta, which satisfies $\delta_{k,k'} = 1$ if $k = k'$ and $\delta_{k,k'} = 0$ otherwise. The number of nodes of degree k in a tree which consists of s nodes, and includes a single hub with b branches is

$$N(K = k) = b\delta_{k,1} + (s - 1 - b)\delta_{k,2} + \delta_{k,b}. \tag{66}$$

The number of nodes of degree k in a tree which consists of s nodes, and takes the form of a chain of h hubs, with a total of b branches distributed according to $\vec{b} = (b_1, b_2, \dots, b_h)$, is

$$N(K = k) = b\delta_{k,1} + (s - h - b)\delta_{k,2} + \sum_{i=2}^{h-1} \delta_{k,b_i+2} + \delta_{k,b_1+1} + \delta_{k,b_h+1}. \tag{67}$$

Consider a tree of topology $\tau = (h, A, \vec{b})$ which consists of s nodes. Such tree includes h hubs, whose degrees in the backbone tree are given by $\vec{a} = (a_1, a_2, \dots, a_h)$ and their branch vector is $\vec{b} = (b_1, b_2, \dots, b_h)$. The number of nodes of degree k is given by

$$N(k|\tau; s) = b\delta_{k,1} + (s - h - b)\delta_{k,2} + \sum_{i=1}^{h} \delta_{k,a_i+b_i}. \tag{68}$$

The degree distribution, $P_{\rm FC}(K = k|\tau; s)$, of trees of topology τ, which consist of s nodes, is given by

$$P_{\rm FC}(K = k|\tau; S = s) = \frac{N(k|\tau; s)}{s}, \tag{69}$$

where $N(k|\tau; s)$ is given by Eq. (68). In the analysis below we use different types of degree distributions. In Table I we summarize these distributions and list the equations from which each one of them can be evaluated.

The degree distribution over all the tree topologies which consist of s nodes is given by

$$P_{\rm FC}(K = k|S = s) = \frac{\sum_{\{\tau|s\}} X(\tau)W(\tau; s)P_{\rm FC}(K = k|\tau; S = s)}{\sum_{\{\tau|s\}} X(\tau)W(\tau; s)}, \tag{70}$$

where $k = 1, 2, \dots, s - 1$, the probabilities $P_{\rm FC}(K = k|\tau; S = s)$ are given by Eq. (69) and the summation is over all cluster topologies which can be constructed from s nodes.

In Table II we present the probabilities $P_{\rm FC}(K=k|S=s)$ for trees of $s=2,3,\dots,10$ nodes. These probabilities are determined by combinatorial considerations **COMMENT:** *Obtained by hand or by a computer enummeration?* and are expressed in terms of constant rational numbers.

Summing up the degree distributions obtained from Eq. Eq. (70) over all the tree topologies which consist of up to s nodes, with suitable weights, we obtain

$$P_{\rm FC}(K=k|S\le s)=\frac{\sum_{s'=2}^{s} s' P_{\rm FC}(S=s') P_{\rm FC}(K=k|S=s')}{\sum_{s'=2}^{s} s' P_{\rm FC}(S=s')}. \tag{71}$$

This equation provides an exact analytical expression for the degree distribution over all tree topologies up to any desired size, s. In Table III we present these expressions for $P_{\rm FC}(K=k|S\le s)$ where $s=2,3,\dots,6$ and $k=1,2,\dots,5$. It turns out that in these expressions the mean degree, c, always appears in the form $\eta=ce^{-c}$.

Expanding the results of Eq. (71) in powers of the small parameter c, we obtain

$$P_{\rm FC}(K=k|S\le s)=\frac{e^{-c}c^k}{(1-e^{-c})k!}\left(1+q_{s,k}c^{s-k}+\dots\right), \tag{72}$$

where $k=1,2,\dots,s-1$ and the coefficients $q_{s,k}$ are rational numbers of order 1.

As expected, as s is increased the degree distribution given by Eq. (72) converges to the asymptotic form given by

$$\pi_{\rm FC}(K=k)=\frac{e^{-c}c^k}{(1-e^{-c})k!}, \tag{73}$$

which is the degree distribution of the whole subcritical ER network, except for the isolated nodes. This convergence to the Poisson degree distribution confirms the validity of the topological expansion and shows that the combinatorial factors were evaluated correctly. In Table IV we present the leading correction terms, $q_{s,k}c^{s-k}$, of Eq. (72), obtained from the topological expansion, for all the tree structures which consist of up to s nodes, where $s=2,3,\dots,10$. Tree structures with up to s nodes support degrees in the range of $k=1,\dots,s-1$.

VII. THE MEAN AND VARIANCE OF THE DEGREE DISTRIBUTION

The moments of the degree distribution provide useful information about the network structure. The first and second moments are of particular importance. The first moment, $\langle K \rangle_{\rm FC}$ provides the mean degree. The width of the distribution is characterized by the variance, ${\rm Var}(K) = \langle K^2 \rangle - \langle K \rangle^2$, where $\langle K^2 \rangle$ is the second moment.

The $n^{\rm th}$ moment of the degree distribution over all trees of topology τ which consist of s nodes can be expressed by

$$\mathbb{E}[K^n|\tau; S=s] = \sum_{k=1}^{s-1} k^n P_{\rm FC}(K=k|\tau; S=s), \quad (74)$$

where $P_{\rm FC}(K=k|\tau; S=s)$ is given by Eq. (69). The $n^{\rm th}$ moment of the degree distribution over all tree topologies which consist of s nodes is given by

$$\mathbb{E}[K^n|S=s] = \frac{\sum_{\{\tau|s\}} X(\tau) W(\tau; s) \mathbb{E}[K|\tau; S=s]}{\sum_{\{\tau|s\}} X(\tau) W(\tau; s)}, \quad (75)$$

where $\mathbb{E}[K^n|\tau; S=s]$ is given by Eq. (74). For the special case of $n=1$, one obtains

$$\mathbb{E}[K|S=s] = 2 - \frac{2}{s}. \quad (76)$$

This result represents a topological invariance and it applies to any tree of s nodes, regardless of its topology, τ. This is due to the fact that any tree of s nodes includes $s-1$ edges and each edge is shared by two nodes. The results for the first two moments, $\mathbb{E}[K|S=s]$ and $\mathbb{E}[K^2|S=s]$, and for the variance ${\rm Var}[K|S=s] = \mathbb{E}[K^2|S=s] - (\mathbb{E}[K|S=s])^2$, for $s=2,3,\dots,10$ are shown in Table II.

The $n^{\rm th}$ moment of the degree distribution over all trees which consist of up to s nodes is given by

$$\mathbb{E}[K^n|S \le s] = \frac{\sum_{s'=2}^{s} s' P_{\rm FC}(S=s') \mathbb{E}[K^n|S=s']}{\sum_{s'=2}^{s} s' P_{\rm FC}(S=s')}, \quad (77)$$

where $P_{\rm FC}(S=s')$ is given by Eq. (25). Performing the summation for a given value of s provides an exact analytical expression for the $n^{\rm th}$ moment of the degree distribution over all tree topologies which consist of up to s nodes. The resulting expressions for the mean degree, $\mathbb{E}[K|S \le s]$, over all trees which consist of up to $s=2,3,\dots,6$ nodes, are presented in Table III. In the limit of large s, the mean degree $\mathbb{E}[K|S \le s]$ converges towards the asymptotic result, which is given by

$$\langle K \rangle_{\rm FC} = \frac{c}{1-e^{-c}}. \tag{78}$$

In Fig. 4 we present the mean degrees, $\mathbb{E}[K|S \le s]$, as a function of c (thin solid lines). The results are shown for all tree topologies of sizes smaller or equal to s, where $s = 2, 3, \dots, 10$ (from bottom to top). The thick solid line shows the asymptotic result, $\langle K \rangle_{\rm FC}$, given by Eq. (78).

Below we derive closed form analytical expressions for the mean degree, $\mathbb{E}[K|S \le s]$, over all tree topologies which consist of up to s nodes. Inserting the expression for $\mathbb{E}[K|S = s]$ from Eq. (76) into equation (77), with $n = 1$, we obtain

$$\mathbb{E}[K|S \le s] = 2 - 2\frac{\sum_{s'=2}^{s} P_{\rm FC}(S = s')}{\sum_{s'=2}^{s} s' P_{\rm FC}(S = s')}. \tag{79}$$

This result can be expressed in the form

$$\mathbb{E}[K|S \le s] = 2 - \frac{2 - c - 2e^{-c} - 2\sum_{s'=s+1}^{\infty} P_{\rm FC}(S = s')}{1 - e^{-c} - \sum_{s'=s+1}^{\infty} s' P_{\rm FC}(S = s')}. \tag{80}$$

Expressing the distribution $P_{\rm FC}(S = s')$ by Eq. (25) we obtain

$$\mathbb{E}[K|S \le s] = 2 - \frac{\sqrt{2\pi}(2 - c - 2e^{-c}) - c^s e^{-(c-1)(s+1)} \Phi\left(ce^{1-c}, \frac{5}{2}, s+1\right)}{\sqrt{2\pi}(1 - e^{-c}) - c^s e^{-(c-1)(s+1)} \Phi\left(ce^{1-c}, \frac{3}{2}, s+1\right)}, \tag{81}$$

where $\Phi(z, s, a)$ is the Hurwitz Lerch Phi transcendent. An alternative approach for the evaluation of $\mathbb{E}[K|S \le s]$ is to go back to Eq. (80) and replace the sums, $\sum_{s'=s+1}^{\infty}$ by integrals of the form $\int_{s+1/2}^{\infty}$. Performing the integrations, we obtain

$$\mathbb{E}[K|S \le s] = 2 - \frac{\sqrt{2\pi} c(2 - c - 2e^{-c}) - (c - 1 - \ln c)^{3/2} \Gamma\left[-\frac{3}{2}, (c - 1 - \ln c)\left(s + \frac{1}{2}\right)\right]}{\sqrt{2\pi} c(1 - e^{-c}) - (c - 1 - \ln c)^{1/2} \Gamma\left[-\frac{1}{2}, (c - 1 - \ln c)\left(s + \frac{1}{2}\right)\right]}, \tag{82}$$

where $\Gamma(s, a)$ is the incomplete Gamma function. This function satisfies

$$\Gamma\left(-\frac{3}{2}, x\right) = \frac{4}{3}\sqrt{\pi}\left[1 - \text{erf}(\sqrt{x})\right] + \frac{2e^{-x}(1 - 2x)}{3x^{3/2}}, \tag{83}$$

and

$$\Gamma\left(-\frac{1}{2},x\right) = -2\sqrt{\pi}\left[1-\mathrm{erf}(\sqrt{x})\right] + \frac{2e^{-x}}{\sqrt{x}}, \tag{84}$$

where $\mathrm{erf}(x)$ is the error function.

For $c = 1$ the Phi transcendent function in Eq. (81) can be replaced by the Hurwitz Zeta function. In this case

$$\mathbb{E}[K|S \le s] = 2 - \frac{\sqrt{2\pi}(2e-4) - 4e\zeta\left(\frac{5}{2}, s+1\right)}{\sqrt{2\pi}(e-1) - e\zeta\left(\frac{3}{2}, s+1\right)}, \tag{85}$$

where $\zeta(s,a)$ is the Hurwitz zeta function. In the limit of large s, one can approximate Eq. (85) by an asymptotic expansion of the form

$$\mathbb{E}[K|S \le s] = \frac{e}{e-1} - \sqrt{\frac{2}{\pi}}\frac{e(e-2)}{(e-1)^2}\frac{1}{\sqrt{s}} - \frac{2}{\pi}\frac{e^2(e-2)}{(e-1)^3}\frac{1}{s} + \mathcal{O}\left(\frac{1}{s^{3/2}}\right). \tag{86}$$

In Fig. 5 we present the mean degree $\mathbb{E}[K|S \le s]$ over all trees of $S \le s$ nodes, as a function of s at the critical value of $c = 1$. The analytical results (circles), obtained from Eq. (85), are in excellent agreement with the exact results of the asymptotic expansion (solid line). The only slight deviations are for $s = 2$ and 3, and they reflect the fact that Eq. (85) is based on the Stirling expansion which becomes accurate for $s \ge 4$. The results of the asymptotic expansion to order $1/\sqrt{s}$ ($\times$), obtained from the first two terms of Eq. (86), exhibit large deviations from the exact results, particularly for small values of s. This means that next order correction should be taken into account, at least for such small values of s. Indeed, an expansion to order $1/s$ obtained by including the third term in Eq. (86), greatly improves the results (+).

Using Eq. (77) one can obtain exact analytical expressions for the second moment of the degree distribution over all trees of size $S \le s$. The results for small trees which consist of up to $s = 2, 3, \dots, 6$ nodes are shown in Table III. In the limit of large s, the second moment $\mathbb{E}[K^2|S \le s]$ converges towards the asymptotic result, which is given by

$$\langle K^2 \rangle_{\mathrm{FC}} = \frac{c(c+1)}{1-e^{-c}}. \tag{87}$$

The variance of the degree distribution over all trees which consist of up to s nodes is given by

$$\mathrm{Var}[K|S \le s] = \mathbb{E}[K^2|S \le s] - (\mathbb{E}[K|S \le s])^2. \tag{88}$$

Using the results presented in Table III for the first and second moments of the degre distributions over small trees of sizes $s = 2, 3, \dots, 5$, we obtain

$$\begin{aligned}
\mathrm{Var}[K|S \le 2] &= 0 \\
\mathrm{Var}[K|S \le 3] &= \frac{2\eta + 2\eta^2}{(2+3\eta)^2} \\
\mathrm{Var}[K|S \le 4] &= \frac{18\eta + 78\eta^2 + 90\eta^3 + 96\eta^4}{(6 + 9\eta + 16\eta^2)^2} \\
\mathrm{Var}[K|S \le 5] &= \frac{288\eta + 1248\eta^2 + 3960\eta^3 + 5016\eta^4 + 6920\eta^5 + 7500\eta^6}{(24 + 36\eta + +64\eta^2 + 125\eta^3)^2}.
\end{aligned} \tag{89}$$

In the limit of large s, the variance $\mathrm{Var}[K|S \le s]$ converges towards the asymptotic result, $\sigma^2_{K,\mathrm{FC}} = \mathrm{Var}(K)$, where $\sigma_{k,\mathrm{FC}}$ is the standard deviation of the degree distribution over all the finite clusters. The asymptotic variance is given by

$$\sigma^2_{K,\mathrm{FC}} = \mathrm{Var}(K) = \frac{c}{1-e^{-c}} - \frac{c^2 e^{-c}}{(1-e^{-c})^2}. \tag{90}$$

VIII. THE DISTRIBUTION OF SHORTEST PATH LENGTHS

In this Section we apply the topological expansion to obtain the DSPL of subcritical ER networks and to express it in terms of the contributions of the different tree topologies. Summing up the contributions for all possible tree topologies supported by up to s nodes, we express the DSPL as a power series in c, and find its asymptotic form in the limit of $N \to \infty$.

For each value of $s = 2, 3, \dots,$ we identify all the tree topologies, τ, supported by s nodes. For each one of these tree topologies, and for $\ell = 1, 2, \dots, s-1$, we calculate the number, $N(L = \ell|\tau; s)$ of pairs of nodes which reside at a distance ℓ from each other. We then sum up these contributions over all the possible ways to assemble s nodes into the given tree topology. Below we describe the enumeration of the shortest paths for a few simple examples of tree topologies.

In a linear chain of s nodes there are $s - \ell$ pairs of nodes at distance ℓ from each other. Therefore,

$$N[L = \ell|\tau = (0, \cdot, 2); s] = \binom{s-\ell}{1} \tag{91}$$

A convenient way to evaluate the number of such pairs is to take a pair of nodes at a distance ℓ from each other and reduce the chain of $\ell+1$ nodes between them into a single node which is marked to distinguish it from the other nodes. This results in a reduced network of $k-\ell$ nodes, one of which is the marked node. At this point, counting the number of pairs of nodes which are at a distance ℓ from each other is equivalent to counting the number of different locations of the marked node in the reduced network. In Fig. 6 we illustrate this procedure for the case of a linear chain of nodes. Since each node in the reduced chain may be the marked node, one concludes that in the original chain there are $s-\ell$ pairs of nodes at a distance ℓ from each other.

For a tree of s nodes which includes a single hub and b branches, the number of pairs of nodes at a distance ℓ from each other is

$$N[L=\ell|\tau=(1,0,b),s] = b\binom{s-\ell}{b} + (\ell-1)\binom{b}{2}\binom{s-\ell}{b-1}. \tag{92}$$

In this case there are many different configurations due to the different ways to distribute the nodes between the b branches. Therefore, we need to sum up the numbers of pairs of nodes at distance $L=\ell$ from each other in all the different configurations. In this calculation we distinguish between pairs of nodes which reside on the same branch and pairs of nodes which reside on different branches. To calculate the number of pairs of nodes residing on the same branch and are at a distance $L=\ell$ apart, we pick one such pair of nodes and reduce the chain of $\ell+1$ nodes between them into a single node. This node is marked in order to keep track of its location. The reduced network now consists of $s-\ell$ nodes. We then evaluate the number of ways to distribute these $s-\ell$ nodes between the b branches **COMMENT:** *Isn't this (partial) double counting, because the number of trees, i.e., ways to distribute nodes for a given topology, is already included in the $W(\tau;s)$ factor which is later multiplied? I would think, that when both nodes are on the same branch, each branch of length l contributes like a single linear chain, i.e., there are $l_b-\ell$ ways = shortest paths, this just has to be summed over all branches ... Or is it included here, because in Eq. 104 its is divided by $W(\tau;s)$? If yes, maybe state that the approach is a bit different "built up" here compared to the degree case Eq. 69.* and the number of ways to place the marked node in its own branch. Essentially, the marked node splits its branch into two parts. This means that the number of possible configurations is equal to the number of possible ways to distribute $s-\ell$ nodes to $b+1$ urns. The first binomial coefficient in Eq. (92) accounts for the number of such distributions.

To calculate the number of pairs of nodes, i and j, which reside on different branches and are at a distance ℓ apart from each other, we first arrange all s nodes in a linear chain. We choose a pair of nodes which are at a distance ℓ from each other and reduce the $\ell+1$ nodes between them into a single node. This results in a reduced chain of $s-\ell$ nodes, one of which is the marked node. We proceed in two stages. In the first stage we consider the two branches on which the nodes i and j reside as a single branch, which now includes the marked node. The binomial coefficient $\binom{s-\ell}{b-1}$ accounts for the number of ways to distribute the nodes into $b-1$ urns and to choose randomly the location of the marked node. In the second stage we randomly choose the location of the hub among the $\ell-1$ nodes between i and j and connect all the end-points of all other $b-2$ branches to this node. Apart from this, there are $\binom{b}{2}$ possible ways to choose the branches on which i and j are located.

The approach presented above can also be used to evaluate the number of pairs of nodes at a distance $L=\ell$ apart, which reside on branches which do not share a hub. In this case one needs to account for the number of possible ways to locate two or more hubs along the segment of $\ell-1$ nodes between i and j. For a tree of s nodes, which includes two hubs, we obtain

$$\begin{aligned} N[L=\ell|\tau=(2,A,\vec{b});s] &= (b_1+b_2+1)\binom{s-\ell}{b_1+b_2+1} \\ &+ (\ell-1)\left[\binom{b_1+1}{2}+\binom{b_2+1}{2}\right]\binom{s-\ell}{b_1+b_2} \\ &+ b_1 b_2\binom{\ell-1}{2}\binom{s-\ell}{b_1+b_2-1}, \end{aligned} \tag{93}$$

where A is given by Eq. (36) and $\vec{b}=(b_1,b_2)$. Generalizing this result to the case of a linear chain of h hubs we obtain

$$
\begin{aligned}
N[L=\ell|\tau=(h,A,\vec{b}),s] &= (b+h-1)\binom{\ell-1}{0}\binom{s-\ell}{b+h-1} \\
&+\left[\binom{b_1+1}{2}+\sum_{i=2}^{h-1}\binom{b_i+2}{2}+\binom{b_h+1}{2}\right]\binom{\ell-1}{1}\binom{k-\ell}{b+h-2} \\
&+\sum_{r=2}^{h-1}\left[b_1(b_{r+1}+1)+\sum_{i=1}^{h-r-1}(b_i+1)(b_{i+r}+1)+(b_{h-r}+1)b_h\right] \\
&\times\binom{\ell-1}{r}\binom{s-\ell}{b+h-r-1}+b_1b_h\binom{\ell-1}{h}\binom{s-\ell}{b-1},
\end{aligned} \tag{94}
$$

where A is an $h \times h$ Toeplitz matrix which satisfies $A_{ij}=1$ if $j=i\pm 1$ and $A_{ij}=0$ otherwise. Similarly, for a tree which consists of a central hub which is surrounded by $h-1$ peripheral hubs, $N(L=\ell|\tau;s)$ is given by

$$
\begin{aligned}
N[L=\ell|\tau=(h,A,\vec{b});s] &= (b+h-1)\binom{\ell-1}{0}\binom{s-\ell}{b+h-1} \\
&+\left[\binom{b_1+h-1}{2}+\sum_{i=2}^{h}\binom{b_i+1}{2}\right]\binom{\ell-1}{1}\binom{s-\ell}{b+h-2} \\
&+(b_1+h-2)\sum_{i=2}^{h}b_i\binom{\ell-1}{2}\binom{s-\ell}{b+h-3} \\
&+\sum_{i=2}^{h}\sum_{j=i+1}^{h}b_ib_j\binom{\ell-1}{3}\binom{s-\ell}{b+h-4},
\end{aligned} \tag{95}
$$

where $A_{1j}=1$ for $j \geq 2$, $A_{i1}=1$ for $i \geq 2$ and $A_{ij}=0$ otherwise.

We will now derive an equation for the number of pairs of nodes at a distance ℓ from each other in any given tree of s nodes, whose structure is given by the topology $\tau=(h,A,\vec{b})$. Such tree includes h hubs, whose degrees are given by the vector

$$
\vec{k}=(k_1,k_2,\ldots,k_h), \tag{96}
$$

where $k_i=a_i+b_i$. For convenience we also define the vector

$$
\vec{k}'=(k_1-1,k_2-1,\ldots,k_h-1). \tag{97}
$$

The hubs form a backbone tree of h nodes, described by the adjacency matrix, A, of dimensions $h \times h$. For any pair of hubs i and j which are connected by an arm (regardless of its length in the complete tree), the matrix element $A_{ij} = 1$, while otherwise $A_{ij} = 0$. From the adjacency matrix, A, one can obtain the $h \times h$ distance matrix, D, of the backbone tree, which consists of the hubs alone. This is a symmetric matrix, whose matrix element D_{ij} is the distance between hub i and hub j on the backbone tree, and the diagonal elements are $D_{ii} = 0$. For the analysis presented below, it is useful to express the distance matrix as a sum of symmetric binary matrices in the form

$$D = D_{\mathbf{1}} + 2D_{\mathbf{2}} + 3D_{\mathbf{3}} + \cdots + (h-1)D_{h-1}, \tag{98}$$

where $(D_\ell)_{ij} = 1$ if $D_{ij} = \ell$ and $(D_\ell)_{ij} = 0$ otherwise. The matrix D_ℓ, $\ell = 1, 2, \dots, h-1$ is called the $\ell^{\rm th}$ order vertex-adjacency matrix [46]. It can be obtained directly from the adjacency matrix, A, by constructing its powers A^1, A^2, ..., A^ℓ. In case that $(A^\ell)_{ij} = 1$, under the condition that $(A^{\ell'})_{ij} = 0$ for all the lower powers of A, namely $\ell' = 1, 2, \dots, \ell-1$, then $(D_\ell)_{ij} = 1$, and otherwise $(D_\ell)_{ij} = 0$.

Each pair of nodes, i and j in the network can be classified according to the number of hubs, ν_{ij}, which reside along the path between them. If i and j reside on the same branch or on the same arm, $\nu_{ij} = 0$. If they reside on different branches or arms which emanate from the same hub, $\nu_{ij} = 1$. In case that i and j reside on branches or arms which do not share a hub, we denote by h_i the hub which is nearest to i along the path to j and by h_j the hub which is nearest to j along the path to i. We denote by D_{ij} the distance between the hubs h_i and h_j on the backbone tree, which consists of the hubs alone. The number of hubs along the shortest path between nodes i and j can be expressed by $\nu_{ij} = D_{ij} + 1$. Thus, ν_{ij} may take values in the range $0 \le \nu_{ij} \le h$.

The number of pairs of nodes which are at a distance ℓ from each other can be expressed in the form

$$N(L = \ell | \tau, s) = \sum_{\nu=0}^{h} N_\nu(L = \ell | \tau, s), \tag{99}$$

where $N_\nu(L = \ell | \tau, s)$ is the number of pairs of nodes, i and j which are at a distance ℓ from each other and along the path between them there are ν hubs.

For pairs of nodes which reside on the same branch or on the same arm, for which $\nu = 0$, we obtain

$$N_0(L=\ell|\tau,s) = (b+h+1)\binom{\ell-1}{0}\binom{s-\ell}{b+h-1}. \quad (100)$$

For pairs of nodes which reside on different branches or arms which emanate from the same hub, for which $\nu = 1$, we obtain

$$N_1(L=\ell|\tau,s) = \binom{\ell-1}{1}\binom{s-\ell}{b+h-2}\sum_{i=1}^{h}\binom{k_i}{2}. \quad (101)$$

For pairs of nodes for which $\nu \geq 2$ we obtain

$$N_\nu(L=\ell|\tau,s) = \frac{1}{2}\binom{\ell-1}{\nu}\binom{s-\ell}{b+h-\nu-1}\sum_{i=1}^{h}\sum_{j=1}^{h} k'_i k'_j \delta_{D_{ij},\nu-1}. \quad (102)$$

Eq. (102) can be written in the form

$$N_\nu(L=\ell|\tau,s) = \frac{1}{2}\binom{\ell-1}{\nu}\binom{s-\ell}{b+h-\nu-1}\vec{k}'^T D_{\nu-1}\vec{k}', \quad (103)$$

where $\vec{k}'^T$ is the transpose of $\vec{k}'$.

The distribution $P_{\rm FC}(L=\ell|\tau;L<\infty,S=s)$, for trees of a given topology, τ, assembled from s nodes, is given by

$$P_{\rm FC}(L=\ell|\tau;L<\infty,S=s) = \frac{N(L=\ell|\tau;s)}{\binom{s}{2}W(\tau;s)}. \quad (104)$$

In the analysis below we use different types of DSPLs. In Table I we summarize these distributions and list the equations from which each one of them can be evaluated.

The DSPL over clusters of all topologies which consist of s nodes is given by

$$P_{\rm FC}(L=\ell|L<\infty,S=s) = \frac{\sum_{\{\tau|s\}} X(\tau)W(\tau;s)P_{\rm FC}(L=\ell|\tau;L<\infty,S=s)}{\sum_{\{\tau|s\}} X(\tau)W(\tau;s)}, \quad (105)$$

COMMENT: *just above it is divided by $W(\tau;s)$ and multiplied here again, so it is somehow irrelevant? If intended, please explain/justify, maybe it is divided above because of double counting avoidance and multiplied here such that it has the same structure as for the degrees?* where the summation is over all cluster topologies which can be constructed from s nodes, In Table V we present the probabilities $P_{\rm FC}(L=\ell|L<\infty,S=s)$ for trees of $s=2,3,\ldots,10$ nodes. These probabilities are determined by combinatorial considerations and are expressed in terms of constant rational numbers.

To obtain the DSPL over all the clusters of sizes $s' \le s$, we sum up the results of Eq. (105) over all these clusters:

$$P_{\rm FC}(L=\ell | L<\infty, S \le s) = \frac{\sum_{s'=2}^{s} \binom{s'}{2} P_{\rm FC}(S=s') P_{\rm FC}(L=\ell|S=s')}{\sum_{s'=2}^{s} \binom{s'}{2} P_{\rm FC}(S=s')}. \tag{106}$$

This equation provides an exact analytical expression for the degree distribution over all tree topologies up to any desired size, s. In Table VI we present these expressions for $P_{\rm FC}(L=\ell|L<\infty, S\le s)$ where $s = 2, 3, \dots, 6$ and $\ell = 1, 2, \dots, 5$. It turns out that in these expressions the mean degree, c, always appears in the form $\eta = ce^{-c}$.

Expanding the results of Eq. (106) as a power series in the small parameter c, we find that

$$P_{\rm FC}(L=\ell|L<\infty, S\le s) = (1-c)c^{\ell-1}\left(1 + r_{s,\ell}c^{s-\ell} + \dots\right), \tag{107}$$

where $\ell = 2, 3, \dots, s-1$ and the coefficient $r_{s,\ell}$ is a rational number of order 1. In Table VII we present the leading finite size correction terms, $r_{s,\ell}c^{s-\ell}$, of Eq. (107), obtained from the topological expansion, for all the tree structures which consist of up to s nodes, where $s = 2, 3, \dots, 10$. Tree structures with up to s nodes support distances in the range of $\ell = 1, \dots, s-1$. In the limit of large s, **because all powers of $c < 1$ will disappear,** these results converge towards the asymptotic form

$$P_{\rm FC}(L=\ell|L<\infty) = (1-c)c^{\ell-1}, \tag{108}$$

which turns out to be the DSPL of the entire subcritical network in the asymptotic limit of $N \to \infty$. In spite of its apparent simplicity, this is a surprising and nontrivial result, which was not anticipated when we embarked on the topological expansion. Eq. (108) is essentially a mean field result. Normally, a mean field result for the DSPL is expected to represent the shell structure around a typical node. However, in this case there is no typical node. The shell structures around each node strongly depends on the size and topology of the cluster on which it resides as well as on its location in the cluster. Only by combining the contributions of all pairs of nodes one obtains the simple expression of Eq. (108).

The DSPL given by Eq. (108) is a conditional distribution, under the condition that the selected pair of nodes reside on the same cluster. In fact, it is a subleading component of the overall DSPL of the network, because in subcritical networks most pairs of nodes reside on different clusters, and are thus at an infinite distance from each other. The overall DSPL can be expressed by

$$P_{\rm FC}(L=\ell) = P_{\rm FC}(L<\infty) P_{\rm FC}(L=\ell | L<\infty), \tag{109}$$

where $P_{\rm FC}(L<\infty)$ is given by Eq. (21). Therefore,

$$P_{\rm FC}(L=\ell) = \frac{c^\ell}{N} \tag{110}$$

and

$$P_{\rm FC}(L=\infty) = 1 - \frac{c}{(1-c)N}. \tag{111}$$

The tail distribution which corresponds to the probability distribution function of Eq. (108) is given by

$$P_{\rm FC}(L>\ell | L<\infty) = c^\ell. \tag{112}$$

In Fig. 7 we present theoretical results for the DSPL of asymptotic ER networks with $c = 0.2, 0.4, 0.6$ and 0.8 (solid lines), obtained from Eq. (108). These results are compared to numerical results for the DSPL (symbols), obtained for networks of size $N = 10^4$ and the same four values of c. We find that the theoretical results are in very good agreement with the numerical results except for small deviations in the large distance tails. These deviations are due to finite size of the simulated networks. The numerical simulations were performed via sampling of 10^4 independent realizations of ER networks of size $N = 10^4$ for each value of c [32]. For each realization we applied the *all pairs shortest paths* algorithm from the LEDA C++ library [47].

In Fig. 8 we present the probability $P_{\rm FC}(L=\ell | L<\infty)$, given by Eq. (108), as a function of the mean degree, c, for $\ell = 1, 2, 5$ and 10. The probability $P_{\rm FC}(L=1 | L<\infty)$ is a monotonically decreasing function of c. This is due to the fact that for very low values of c most of the clusters consisting of two or more nodes are dimers and their fraction decreases as c is increased. For $\ell \geq 2$, the probability $P_{\rm FC}(L=\ell | L<\infty)$ vanishes at $c = 0$ and $c = 1$. It increases for low values of c, reaches a peak and then starts to decrease. For each value of $\ell \geq 2$, the peak of $P_{\rm FC}(L=\ell | L<\infty)$ is located at $c = 1 - 1/\ell$, reflecting the appearance of longer paths as c is increased.

It is also interesting to consider the conditional probabilities $P_{\rm FC}(L=\ell | L<\infty, K=k)$ and $P_{\rm FC}(L=\ell | L<\infty, K=k, K'=k')$, between random pairs of nodes which reside on the same finite cluster, under the condition that the degrees of one or both nodes are specified, respectively. In supercritical networks, the paths between nodes of high degrees tend to shorter than between nodes of low degrees. This is due to the fact that higher degrees open more paths between the nodes, increasing the probability of short paths to emerge. The situation in subcritical networks is completely different. Any pair of nodes, i and j, which reside on the same cluster are connected by a single path. Such path goes through one neighbor of i and one neighbor of j. Therefore, the statistics of the path lengths between pairs of nodes which reside on the same cluster in subcritical ER networks do not depend on the degrees of these nodes. As a result, the conditional DSPLs satisfy

$$P_{\rm FC}(L=\ell | L<\infty, K=k) = (1-c)c^{\ell-1}, \tag{113}$$

regardless of the value of k, and

$$P_{\rm FC}(L=\ell | L<\infty, K=k, K'=k') = (1-c)c^{\ell-1}, \tag{114}$$

regardless of the values of k and k'. It is worth pointing out, however, that the probability that a random node of a specified degree, k, and another random node of an unspecified degree reside on the same cluster is dependent on the degree, k. Using the results of Appendix A it can be shown that

$$P_{\rm FC}(L<\infty | K=k) = \frac{k}{(1-c)N}. \tag{115}$$

Similarly, it can be shown that the probability that a random node of degree k and another random node of degree k' reside on the same cluster is given by

$$P_{\rm FC}(L<\infty | K=k, K'=k') = \frac{kk'}{c(1-c)N}. \tag{116}$$

The DSPL of Eq. (108) applies not only for subcritical ER networks but also for the finite clusters of supercritical ER networks. According to the duality relations, given a supercritical ER network of N nodes and mean degree $c>1$, the subnetwork which consists of the finite clusters is a subcritical ER network of size

$$N' = Nf(c), \tag{117}$$

and mean degree

$$c' = cf(c), \tag{118}$$

where $f(c) = -\mathcal{W}(-ce^{-c})/c$ is the fraction of nodes in the supercritical network which reside on the finite clusters and $c' < 1$. In Fig. 9 we present numerical results for the DSPL of the finite clusters of a supercritical ER network of $N = 10^4$ nodes and $c = 1.547$ (circles). The results are found to be in very good agreement with numerical results for its dual network, which consists of $N' = 3882$ nodes and $c' = 0.6$ ($\times$) and with the analytical results for an asymptotic subcritical ER network with $c = 0.6$ (solid line), obtained from Eq. (108).

IX. THE MEAN AND VARIANCE OF THE DSPL

The moments of the DSPL provide useful information about the large scale structure of the network. The first and second moments are of particular importance. The first moment, $\langle L \rangle_{\rm FC}$ provides the mean distance. The width of the DSPL is characterized by the variance, $\mathrm{Var}(L) = \langle L^2 \rangle - \langle L \rangle^2$, where $\langle L^2 \rangle$ is the second moment.

The $n^{\rm th}$ moment of the DSPL over all trees of size s and topology τ is given by

$$\mathbb{E}[L^n|\tau; L < \infty, S = s] = \sum_{\ell=1}^{s-1} \ell P_{\rm FC}(L = \ell|\tau; L < \infty, S = s), \tag{119}$$

where $P_{\rm FC}(L = \ell|L < \infty, S = s)$ is given by Eq. (105). **COMMENT:** *But in the equation you use $P_{\rm FC}(L = \ell|\tau; L < \infty, S = s)$, probably this is meant here.* The $n^{\rm th}$ moment of the DSPL over trees of all topologies, which consist of s nodes is given by

$$\mathbb{E}[L^n|S = s] = \frac{\sum_{\{\tau|s\}} X(\tau)W(\tau; s)\mathbb{E}[L|\tau; L < \infty, S = s]}{\sum_{\{\tau|s\}} X(\tau)W(\tau; s)}, \tag{120}$$

where $\mathbb{E}[L|\tau; L < \infty, S = s]$ is given by Eq. (119).

The results for the first two moments, $\mathbb{E}[L|S = s]$ and $\mathbb{E}[L^2|S = s]$, and for the variance $\mathrm{Var}[L|S = s] = \mathbb{E}[L^2|S = s] - (\mathbb{E}[L|S = s])^2$, for $s = 2, 3, \dots, 10$ are shown in Table V.

The $n^{\rm th}$ moment of the DSPL over all trees which consist of up to s nodes is given by

$$\mathbb{E}[L^n|S \le s] = \frac{\sum_{s'=2}^{s} \binom{s'}{2} P_{\rm FC}(S=s') \mathbb{E}[L^n|S=s']}{\sum_{s'=2}^{s} \binom{s'}{2} P_{\rm FC}(S=s')}, \tag{121}$$

where $P_{\rm FC}(S=s')$ is given by Eq. (25) and $\mathbb{E}[L^n|S=s']$ is given by Eq. (120). Performing the summation over all tree topologies up to size s provides exact analytical expressions for the moments of the DSPL over these trees. The results for $\mathbb{E}[L|S \le s']$ and $\mathbb{E}[L^2|S \le s']$ For small trees of sizes $s = 2, 3, \dots, 6$ are shown in Table VI. In the limit of large s, the mean distance $\mathbb{E}[L|S \le s]$ converges towards the asymptotic result, which is given by

$$\langle L \rangle_{\rm FC} = \frac{1}{1-c}. \tag{122}$$

In Fig. 10 we present the mean distances $\mathbb{E}[L|S \le s]$ (solid lines) over all tree topologies of sizes smaller or equal to s, as a function of c, for $s = 2, 3, \dots, 10$ (from bottom to top, respectively). The thick solid line shows the asymptotic result, $\langle L \rangle_{\rm FC}$, given by Eq. (122). **Obviously, with increasing value s of the tree size, $\mathbb{E}[L|S \le s]$ approaches the asymptotic result. As expected, for small values of c the convergence is fast, but about beyond $c = 0.5$, approaching the percolation transiton, the (diverging) asymptotic result is far away and the convergence is slow. For this reason, we compare below with computer simulations again.**

Using Eq. (121) one can obtain exact analytical expressions for the second moment of the DSPL over all trees of size $S \le s$. The results for small trees which consist of up to $s = 2, 3, \dots, 6$ nodes are shown in Table VI. In the limit of large s, the second moment $\mathbb{E}[L^2|S \le s]$ converges towards the asymptotic result, which is given by

$$\langle L^2 \rangle_{\rm FC} = \frac{1+c}{(1-c)^2}. \tag{123}$$

The variance of the DSPL over all trees which consist of up to s nodes is given by

$$\mathrm{Var}[L|S \le s] = \mathbb{E}[L^2|S \le s] - (\mathbb{E}[L|S \le s])^2. \tag{124}$$

Using the results presented in Table VI for the first and second moments of the degre distributions over small trees of sizes $s = 2, 3, \dots, 5$, we obtain

$$\begin{aligned}
\mathrm{Var}[L|S \le 2] &= 0 \\
\mathrm{Var}[L|S \le 3] &= \frac{\eta + 2\eta^2}{(1+3\eta)^2} \\
\mathrm{Var}[L|S \le 4] &= \frac{\eta + 9\eta^2 + 19\eta^3 + 31\eta^4}{\left(1+3\eta+8\eta^2\right)^2} \\
\mathrm{Var}[L|S \le 5] &= \frac{36\eta + 324\eta^2 + 1854\eta^3 + 4044\eta^4 + 7950\eta^5 + 12054\eta^6}{\left(6+18\eta+48\eta^2+125\eta^3\right)^2}.
\end{aligned} \tag{125}$$

In the limit of large s, the variance $\mathrm{Var}[L|S \le s]$ converges towards the asymptotic result, $\sigma^2_{L,\mathrm{FC}} = \mathrm{Var}(L)$, where $\sigma_{L,\mathrm{FC}}$ is the standard deviation of the DSPL over all the finite clusters. The asymptotic variance is given by

$$\mathrm{Var}(L) = \langle L^2 \rangle_{\mathrm{FC}} - \langle L \rangle^2_{\mathrm{FC}}. \tag{126}$$

Using Eqs. (122) and (123) we find that

$$\sigma^2_{L,\mathrm{FC}} = \mathrm{Var}(L) = \frac{c}{(1-c)^2}. \tag{127}$$

In Fig. 11 we present the mean distance, $\langle L \rangle_{\mathrm{FC}}$ (a), and the standard deviation $\sigma_{L,\mathrm{FC}}$ (b), vs. the mean degree, c. The theoretical results (solid lines) correspond to the asymptotic limit. The numerical results, obtained for $N = 10^2$ (+), 10^3 ($\times$) and 10^4 ($\circ$) are found to converge towards the theoretical results as the network size is increased.

X. SUMMARY AND DISCUSSION

We have developed an analytical approach for the calculation of the DSPL in random networks which are fragmented into isolated clusters, which is based on a topological expansion. Applying it to subcritical ER networks, which consist of finite tree components, we derived equations for the DSPL over all tree topologies of up to s nodes. Taking the large s limit, we obtained an exact asymptotic formula for the DSPL over all pairs of nodes which reside on the same cluster, which takes the form $P_{FC}(L = \ell | L < \infty) = (1-c)c^{\ell-1}$. The mean path length was found to be $\langle L \rangle_{\mathrm{FC}} = 1/(1-c)$. As the percolation threshold is approached from below, at $c \to 1^-$, the mean distance diverges as $\langle L \rangle_{\mathrm{FC}} \sim (1-c)^{-\alpha}$, where the exponent $\alpha = 1$. Using the duality relations between a subcritical ER network and the finite clusters in a corresponding supercritical ER network, one can show that the same exponent, $\alpha = 1$, appears also above the transition.

Apart from the shortest path length, random networks exhibit other distance measures such as the resistance distance [48–50]. The resistance distance, r_{ij}, between a pair of nodes, i and j is the electrical resistance between them under conditions in which each edge in the network represents a resistor of 1 Ohm. Unlike the shortest path length, the resistance distance depends on all the paths between i and j, which often merge and split along the way. It can be evaluated using the standard rules under which the total resistance of resistors connected in series is the sum of their individual resistance values, while the total resistance of resistors connected in parallel is the reciprocal of the sum of the reciprocals of the individual resistance values. It was shown that the resistance distance between nodes i and j in a network can be decomposed in terms of the eigenvalues and eigenvectors of the normalized Laplacian matrix of the network [51, 52]. In order to utilize this result for the calculation of the full distribution of resistance distances, $P(R = r)$, in an ensemble of supercritical ER networks, one will need to obtain the full statistics of the spectral properties of the Laplacian matrix over the ensemble, which is expected to be a difficult task. For subcritical ER networks the situation is simpler. Since the finite clusters in subcritical networks are trees, the shortest path between a pair of nodes, i and j, is in fact the only path between them. As a result, the resistance distance between i and j is equal to the shortest path length between them. This means that the results presented in this paper provide not only the distribution of shortest path lengths in subcritical ER networks but also the distribution of resistance distances in these networks, which is given by

$$P_{\rm FC}(R = r | R < \infty) = (1 - c)c^{r-1}, \tag{128}$$

where r takes integer values.

Another distance measure between nodes in random networks is the mean first passage time, t_{ij}, of a random walk (RW) starting from node i and reaching node j [53, 54]. Unlike the shortest path length, the mean first passage time is not symmetric, namely $t_{ij} \neq t_{ji}$. Since a RW may wander through side branches, the mean first passage time cannot be shorter than the shortest path, namely $t_{ij} \geq \ell_{ij}$. However, apart from this inequality, there is no simple way to connect between these two quantities. **Therefore, numerical simulations will be suitable here. Using specific large-deviation algorithms [55, 56], it is possible, in**

principle, to sample the distributions over its full support, i.e. down to very small probabilities like 10^{-100}. Such approaches have been already applied to obtain distributions of several properties of random graphs, e.g., the distribution of the number of components [57], the distribution of the size of the largest component [59], the distribution of the 2-core size [58] or the distribution of the diameter [16].

Unlike RWs which would eventually visit all the nodes in the cluster on which they reside, the paths of self avoiding walks (SAWs) terminate once they enter a leaf node [60]. Therefore, an SAW starting from node i does not necessarily reach node j even if they reside on the same cluster. However, in case it reaches node j its first passage time is equal to the shortest path length between i and j. Therefore, the distribution of first passage times, $P_{\rm SAW}(T = t)|T < \infty)$, of SAWs between pairs of nodes which reside on the same cluster satisfies $P_{\rm SAW}(T = t|T < \infty) = P_{\rm FC}(L = \ell|L < \infty)$.

The DSPL of subcritical ER networks is also relevant to the study of epidemic spreading on supercritical ER networks. Consider a supercritical ER network with mean degree $c > 1$. An epidemic starts from a random node, i, and propagates through the shell structure around this node. The time is discrete, so each node that is infected at time t may infect each one of its neighbors at time $t + 1$, with probability p'. The node that was infected at time t recovers at time $t + 1$ and becomes immune.

The expectation value of the number of nodes infected by node i in the first time step is given by $c' = cp'$. In case that $c' < 1$, the statistical properties of the clusters formed by such epidemic are similar to the statistical properties of the tree components in a subcritical network with mean degree c'. More precisely, the size distribution of clusters formed by the epidemic follows the distribution $P_{\rm FC}(S = s)$ of cluster sizes on which a random node resides, given by Eq. (26). This property represents some kind of invariance, the distribution of epidemic sizes depends only on the value of the product $c' = cp'$ rather than the values of c and p' alone. The DSPL, $P_{\rm FC}(L = \ell|L < \infty)$ represents the temporal propagation of a typical epidemic namely the probability that a node which was infected by an epidemic got infected *ell* time steps after the epidemic started.

Since epedemic spreading and many other real-world processes take place on networks different from ER random networks, it will be interesting to apply the topological expansion presented here to the analysis of the DSPL in subcritical configuration model networks and other complex networks. Using extreme value statistics it may be possible to obtain analytical results for the distributions of radii and diameters over all the tree topologies. For networks which satisfy duality relations, it will be possible to obtain the DSPL on the finite clusters in the supercritical regime. Combining the results with the DSPL on the giant cluster will yield the overall DSPL of the supercritical network. The detailed understanding of the DSPL in terms of the topological expansion is expected to be useful in the study of dynamical processes such as epidemic spreading.

Appendix A: Evaluation of the probability $P_{\rm FC}(L < \infty)$

The number of pairs of nodes which reside on the same cluster is given by

$$\mathcal{L}(N,c) = \sum_{s \ge 1} \binom{s}{2} T_s. \tag{A1}$$

In order to evaluate this sum we use properties of the Lambert W function, denoted by $\mathcal{W}(x)$ [37]. In particular, we use the implicit definition (Eq. 4.13.1 in Ref. [37]):

$$\mathcal{W}(x) = x e^{-\mathcal{W}(x)}. \tag{A2}$$

We also use the series expansion (Eq. 4.13.5 in Ref. [37]):

$$\mathcal{W}(x) = -\sum_{s=1}^{\infty} \frac{s^{s-2}}{s!} s(-x)^s. \tag{A3}$$

Using the series expansion of Eq. (A3) it can be shown that

$$\sum_{s=1}^{\infty} \binom{s}{2} \frac{s^{s-2}}{s!} (-x)^s = \frac{\mathcal{W}(x) - x\mathcal{W}'(x)}{2}. \tag{A4}$$

Plugging in Eq. 4.13.4 of Ref. [37], which takes the form

$$\frac{d\mathcal{W}(x)}{dx} = \frac{\mathcal{W}(x)}{x\left[1+\mathcal{W}(x)\right]}, \tag{A5}$$

we obtain

$$\sum_{s=1}^{\infty} \binom{s}{2} \frac{s^{s-2}}{s!} (-x)^s = \frac{[\mathcal{W}(x)]^2}{2[1+\mathcal{W}(x)]}. \tag{A6}$$

Plugging in $x = -ce^{-c}$, **multiplying with N/c and using the respresentation of T_s^N in Eq. (16), we obtain for the left-hand side**

$$\frac{N}{c}\sum_{s=1}^{\infty}\binom{s}{2}\frac{s^{s-2}}{s!}(ce^{-c})^s = \sum_{s=1}^{\infty}\binom{s}{2}N\frac{s^{s-2}c^{s-1}e^{-cs}}{s!} = \sum_{s=1}^{\infty}\binom{s}{2}T_s^N\,, \tag{A7}$$

which is the quantity we want, for finite values of N. Therefore, we obtain

$$\mathcal{L}(N,c) = \left(\frac{N}{2c}\right)\frac{[\mathcal{W}(-ce^{-c})]^2}{1+\mathcal{W}(-ce^{-c})}. \tag{A8}$$

For $0 < c < 1$ it can be shown that $\mathcal{W}(-ce^{-c}) = -c$, and thus

$$\mathcal{L}(N,c) = \frac{Nc}{2(1-c)}. \tag{A9}$$

As a result, in the asymptotic limit, $N \to \infty$, the probability that two randomly chosen nodes in the network reside on the same cluster is given by

$$P_{\mathrm{FC}}(L < \infty) = \frac{\mathcal{L}(N,c)}{\binom{N}{2}} = \frac{c}{(1-c)N}. \tag{A10}$$

COMMENT: *Shouldn't it be* $\frac{c}{(1-c)(N-1)}$ *which is anyway almost the same?*

[1] S. Havlin and R. Cohen, *Complex Networks: Structure, Robustness and Function* (Cambridge University Press, New York, 2010).

[2] M.E.J. Newman, *Networks: an Introduction* (Oxford University Press, Oxford, 2010).

[3] E. Estrada, *The structure of complex networks: Theory and applications* (Oxford University Press, Oxford, 2011).

[4] P.A.C. Duijn, V. Kashirin and P.M.A and Sloot, *Sci. Rep.*, **4**, 4238 (2014).

[5] J. Shao, S.V. Buldyrev, R. Cohen, M. Kitsak, S. Havlin and H.E. Stanley, *Europhys. Lett.*, **84** 48004 (2008).

[6] J. Shao, S.V. Buldyrev, L.A. Braunstein, S. Havlin and H.E. Stanley *Phys. Rev. E*, **80** 036105 (2009).

[7] P. Erdős and Rényi, *Publ. Math. Debrecen*, **6** 290 (1959); *Publ. Math. Inst. Hungar. Acad. Sci.*, **5** 17 (1960); *Bull. Inst. Internat. Statist*, **38** 343 (1961).

[8] B. Bollobas, *Random Graphs, Second Edition* (Academic Press, London, 2001).

[9] R. Durrett, *Random Graph Dynamics* (Cambridge University Press, Cambridge, 2007).

[10] F. Chung and L. Lu, *Proc. Nat. Acad. Sci. USA* **99** 15879 (2002); *Internet Mathematics*, **1** 91 (2004).

[11] R. van der Hofstad, G. Hooghiemstra and P. Van Mieghem, *Rand. Struct. Alg.*, **27** 76 (2005).

[12] B. Bollobas, S. Janson and O. Riordan, *Rand. Struct. Alg.*, **31**, 3 (2007).

[13] M.E.J. Newman, S.H. Strogatz and D.J. Watts, *Phys. Rev. E*, **64** 026118 (2001).

[14] A. Fronczak, P. Fronczak and J.A. Holyst, *Phys. Rev. E*, **70** 056110 (2004).

[15] T. Łuczak, *Rand. Struct. Alg.*, **13** 485 (1998)

[16] A.K. Hartmann and M. Mézard, arXiv:1710.05680 (2017).

[17] S.N. Dorogotsev, J.F.F. Mendes and A.N. Samukhin, *Nuclear Physics B*, **653** 307 (2003).

[18] V.D. Blondel, J.-L. Guillaume, J.M. Hendrickx and R.M. Jungers, *Phys. Rev. E*, **76** 066101 (2007).

[19] R. van der Hofstad, G. Hooghiemstra and D. Znamenski, *Elect. J. Prob.*, **12** 703 (2007).

[20] H. van der Esker, R. van der Hofstad and G. Hooghiemstra, *J. Stat. Phys.*, **133** 169 (2008).

[21] E. Katzav, M. Nitzan, D. ben-Avraham, P.L. Krapivsky, R. Kühn, N. Ross and O. Biham, *EPL*, **111** 26006 (2015).

[22] M. Nitzan, E. Katzav, R. Kühn and O. Biham, *Phys. Rev. E*, **93** 062309 (2016).

[23] S. Melnik and J.P. Gleeson, arXiv:1604.05521.

[24] B. Gompertz, *Philosophical Trans. R. Soc. London A* **115**, 513 (1825)

[25] R. Pastor-Satorras, C. Castellano, P.V. Mieghem and A. Vespignani, *Rev. Mod. Phys.*, **87** 925 (2015).

[26] A. Barrat, M. Barthélemy and A. Vespignani, *Dynamical processes on complex networks* (Cambridge University Press, New York, 2008).

[27] S. Shao, X. Huang, H.E. Stanley and S. Havlin, *New J. Phys.*, **17** 023049 (2015).

[28] L. Giot et al., *Science* **302** 1727 (2003).

[29] H. Bonneau, A. Hassid, O. Biham, R. Kühn and E. Katzav, *Phys. Rev. E* **96**, 062307 (2017).

[30] A. Goldental, R. Vardi, S. Sardi, P. Sabo and I. Kanter, *Frontiers in Neural Circuits*, **9** 65 (2015).

[31] A. Goldental, H. Uzan, S. Sardi and I. Kanter, *Scientific Reports* **7**, 2700 (2017).

[32] A.K. Hartmann, *Big Practical Guide to Computer Simulations* (World scientific, Singapore, 2015).

[33] R. Cohen and S. Havlin, *Phys. Rev. Lett.*, **90** 058701 (2003).

[34] M. Molloy and A. Reed, *Random Structures and Algorithms* **6**, 161-180 (1995).

[35] M. Molloy and A. Reed, *Combin., Prob. and Comp.* **7**, 295-305 (1998).

[36] H. Klein-Hennig and A.K. Hartmann, Phys. Rev. E **85**, 026101 (2012)

[37] F.W.J. Olver et al., *NIST Handbook of Mathematical Functions* (Cambridge University Press, Cambridge, 2010).

[38] I. Tishby, O. Biham, E. Katzav and R. Kúhn, submitted for publication.

[39] A. Cayley, *Quart. J. Pure Appl. Math.* **23**, 376 (1889).

[40] D. Stauffer D. and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London, 2003).

[41] P. Steinback, *Field guide to simple graphs, Volume 3: The book of trees* (Design Lab, Albouquerque, 1990).

[42] R.A. Wright, B. Richmond, A. Odlyzko and B.D. McKay, *SIAM J. Comput.* **15**, 540 (1986).

[43] T. Beyer and S.M. Hedetniemi, *SIAM J. Comput.* **9**, 706 (1980).

[44] G. Royle, *Trees On Up to 20 Vertices*, http://school.maths.uwa.edu.au/ ∼gordon/remote/graphs/#trees.

[45] R.P. Stanley, *Enumerative combinatorics, Volume I* (Cambridge University Press, Cambride, 2011).

[46] D. Janežič, A. Miličević, S. Nikolić and N. Trinajstić, *Graph-theoretical matrices in chemistry* (CRC Press, Boca Raton, 2015).

[47] K. Mehlhorn and S. Näher, *The LEDA Platform of Combinatorial and Geometric Computing* (Cambridge University Press, Cambridge) 1999.

[48] M.M. Deza and E. Deza, *Encyclopedia of distances*, (Springer, Berlin, 2009).

[49] B. Derrida and J. Vannimenus, *J. Phys. A* **15**, L557 (1982).

[50] R.B. Bapat, *Graphs and matrices* (Springer, London, 2014).

[51] H. Chen and F. Zhang, *Disc. Appl. Math.* **155**, 654 (2007).

[52] Y. Yang and D.J. Klein, *Disc. Appl. Math.* **161**, 2702 (2013).

[53] V. Sood, S. Redner and D. ben-Avraham, *J. Phys. A* **38**, 109 (2005).

[54] C. De Bacco, S.N. Majumdar and P. Sollich, *J. Phys. A* **48**, 205004 (2015).

[55] A.K. Hartmann, Phys. Rev. E **65**, 056102 (2002).

[56] A.K. Hartmann, Phys. Rev. E 89, 052103 (2014).

[57] A. Engel, R. Monasson, and A.K. Hartmann, J. Stat. Phys. 117, 387 (2004).

[58] A.K. Hartmann, Eur. Phys. J. Special Topics 226, 567 (2017).

[59] A.K. Hartmann, Eur. Phys. J. B **84**, 627 (2011).

[60] I. Tishby, O. Biham and E. Katzav, *J. Phys. A* **50**, 115001 (2016).

[61] W. Goddard and O.R. Oellermann, Distance in Graphs, in *Structural Analysis of Complex Networks*, Edited by M. Dehmer, Page 49 (Springer, Dordrecht, 2011).

TABLE I. The different degree distributions and DSPLs for the finite clusters (FC) of subcritical ER networks and the equations which are used to evaluate them.

Distribution	**Equation**	**Description**
$P_{\rm FC}(K=k\|\tau;S=s)$	**Eq. (69)**	The degree distribution over all trees of s nodes and topology τ
$P_{\rm FC}(K=k\|S=s)$	**Eq. (70)**	The degree distribution over all trees of s nodes
$P_{\rm FC}(K=k\|S\le s)$	**Eq. (71)**	The degree distribution over all trees of up to s nodes
$P_{\rm FC}(K=k)$	**Eq. (73)**	The degree distribution over all trees
$\mathbb{E}[K\|\tau;S=s]$	**Eq. (74)**	The mean degree over all trees of s nodes and topology τ
$\mathbb{E}[K\|S=s]$	**Eq. (75)**	The mean degree over all trees of s nodes
$\mathbb{E}[K\|S\le s]$	**Eq. (77)**	The mean degree over all trees of up to s nodes
$\langle K\rangle_{\rm FC}$	**Eq. (78)**	The mean degree over all trees
$P_{\rm FC}(L=\ell\|\tau;L<\infty,S=s)$	**Eq. (104)**	The DSPL over all trees of s nodes and topology τ
$P_{\rm FC}(L=\ell\|L<\infty,S=s)$	**Eq. (105)**	The DSPL over all trees of s nodes
$P_{\rm FC}(L=\ell\|L<\infty,S\le s)$	**Eq. (106)**	The DSPL over all trees of up to s nodes
$P_{\rm FC}(L=\ell\|L<\infty)$	**Eq. (108)**	The DSPL over all trees
$\mathbb{E}[L\|\tau;S=s]$	**Eq. (119)**	The mean distance over all trees of s nodes and topology τ
$\mathbb{E}[L\|S=s]$	**Eq. (120)**	The mean distance over all trees of s nodes
$\mathbb{E}[L\|S\le s]$	**Eq. (121)**	The mean distance over all trees of up to s nodes
$\langle L\rangle_{\rm FC}$	**Eq. (122)**	The mean distance over all trees

TABLE II. The probabilities $P_{\mathrm{FC}}(K=k|S=s)$ that a random node on a random cluster of size s in a subcritical ER network will have a degree k for small tree cluster of $s=2,3,\dots,10$ nodes.

	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$	$s=9$	$s=10$
$P_{\mathrm{FC}}(K=1\|S=s)=$	1	$\frac{2}{3}$	$\frac{9}{16}$	$\frac{64}{125}$	$\frac{625}{1296}$	$\frac{7776}{16807}$	$\frac{117649}{262144}$	$\frac{2097152}{4782969}$	$\frac{43046721}{100000000}$
$P_{\mathrm{FC}}(K=2\|S=s)=$		$\frac{1}{3}$	$\frac{3}{8}$	$\frac{48}{125}$	$\frac{125}{324}$	$\frac{6480}{16807}$	$\frac{50421}{131072}$	$\frac{1835008}{4782969}$	$\frac{4782969}{12500000}$
$P_{\mathrm{FC}}(K=3\|S=s)=$			$\frac{1}{16}$	$\frac{12}{125}$	$\frac{25}{216}$	$\frac{2160}{16807}$	$\frac{36015}{262144}$	$\frac{229376}{1594323}$	$\frac{3720087}{25000000}$
$P_{\mathrm{FC}}(K=4\|S=s)=$				$\frac{1}{125}$	$\frac{5}{324}$	$\frac{360}{16807}$	$\frac{1715}{65536}$	$\frac{143360}{4782969}$	$\frac{413343}{12500000}$
$P_{\mathrm{FC}}(K=5\|S=s)=$					$\frac{1}{1296}$	$\frac{30}{16807}$	$\frac{735}{262144}$	$\frac{17920}{4782969}$	$\frac{45927}{10000000}$
$P_{\mathrm{FC}}(K=6\|S=s)=$						$\frac{1}{16807}$	$\frac{21}{131072}$	$\frac{448}{1594323}$	$\frac{5103}{12500000}$
$P_{\mathrm{FC}}(K=7\|S=s)=$							$\frac{1}{262144}$	$\frac{56}{4782969}$	$\frac{567}{25000000}$
$P_{\mathrm{FC}}(K=8\|S=s)=$								$\frac{1}{4782969}$	$\frac{9}{12500000}$
$P_{\mathrm{FC}}(K=9\|S=s)=$									$\frac{1}{100000000}$
$\mathbb{E}[K\|S=s]=$	1	$\frac{4}{3}$	$\frac{3}{2}$	$\frac{8}{5}$	$\frac{5}{3}$	$\frac{12}{7}$	$\frac{7}{4}$	$\frac{16}{9}$	$\frac{9}{5}$
$\mathbb{E}[K^2\|S=s]=$	1	2	$\frac{21}{8}$	$\frac{76}{25}$	$\frac{10}{3}$	$\frac{174}{49}$	$\frac{119}{32}$	$\frac{104}{27}$	$\frac{99}{25}$
$\mathrm{Var}[K\|S=s]=$	0	$\frac{2}{9}$	$\frac{3}{8}$	$\frac{12}{25}$	$\frac{5}{9}$	$\frac{30}{49}$	$\frac{21}{32}$	$\frac{56}{81}$	$\frac{18}{25}$

TABLE III. The probabilities $P_{\mathrm{FC}}(K=k|S\leq s)$ that a random node on a random cluster of size $S\leq s$ in a subcritical ER network will have a degree k for small tree cluster of $s=2,3,\ldots,10$ nodes.

	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$
$P_{\mathrm{FC}}(K=1\|S\leq s)=$	1	$\frac{2+2\eta}{2+3\eta}$	$\frac{6+6\eta+9\eta^2}{6+9\eta+16\eta^2}$	$\frac{24+24\eta+36\eta^2+64\eta^3}{24+36\eta+64\eta^2+125\eta^3}$	$\frac{120+120\eta+180\eta^2+320\eta^3+625\eta^4}{120+180\eta+320\eta^2+625\eta^3+1296\eta^4}$
$P_{\mathrm{FC}}(K=2\|S\leq s)=$		$\frac{\eta}{2+3\eta}$	$\frac{3\eta+6\eta^2}{6+9\eta+16\eta^2}$	$\frac{12\eta+24\eta^2+48\eta^3}{24+36\eta+64\eta^2+125\eta^3}$	$\frac{60\eta+120\eta^2+240\eta^3+500\eta^4}{120+180\eta+320\eta^2+625\eta^3+1296\eta^4}$
$P_{\mathrm{FC}}(K=3\|S\leq s)=$			$\frac{\eta^2}{6+9\eta+16\eta^2}$	$\frac{4\eta^2+12\eta^3}{24+36\eta+64\eta^2+125\eta^3}$	$\frac{20\eta^2+60\eta^3+150\eta^4}{120+180\eta+320\eta^2+625\eta^3+1296\eta^4}$
$P_{\mathrm{FC}}(K=4\|S\leq s)=$				$\frac{\eta^3}{24+36\eta+64\eta^2+125\eta^3}$	$\frac{5\eta^3+20\eta^4}{120+180\eta+320\eta^2+625\eta^3+1296\eta^4}$
$P_{\mathrm{FC}}(K=5\|S\leq s)=$					$\frac{\eta^4}{120+180\eta+320\eta^2+625\eta^3+1296\eta^4}$
$\mathbb{E}[K\|S\leq 2]=$	1	$\frac{2+4\eta}{2+3\eta}$	$\frac{6+12\eta+24\eta^2}{6+9\eta+16\eta^2}$	$\frac{24+48\eta+96\eta^2+200\eta^3}{24+36\eta+64\eta^2+125\eta^3}$	$\frac{120+240\eta+480\eta^2+1000\eta^3+2160\eta^4}{120+180\eta+320\eta^2+625\eta^3+1296\eta^4}$
$\mathbb{E}[K^2\|S\leq s]=$	1	$\frac{2+6\eta}{2+3\eta}$	$\frac{6+18\eta+42\eta^2}{6+9\eta+16\eta^2}$	$\frac{24+72\eta+168\eta^2+380\eta^3}{24+36\eta+64\eta^2+125\eta^3}$	$\frac{120+360\eta+840\eta^2+1900\eta^3+4320\eta^4}{120+180\eta+320\eta^2+625\eta^3+1296\eta^4}$

TABLE IV. The leading finite size correction terms, $q_{s,k}c^{s-k}$ of Eq. (72) for the degree distribution over all the tree topologies with up to s nodes. The distribution $\pi_{\rm FC}(K=k)$, given by Eq. (73), is the degree distribution over the entire subcritical network, except for the isolated nodes. As s is increased the correction decreases as c^{s-1} and $P_{\rm FC}(K=k|S\le s)$ converges towards $\pi_{\rm FC}(K=k)$.

	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$	$s=9$	$s=10$
$\frac{P_{\rm FC}(K=1\|S\le s)}{\pi_{\rm FC}(K=1)}-1=$	$\frac{1}{2}c$	$\frac{7}{6}c^2$	$\frac{61}{24}c^3$	$\frac{671}{120}c^4$	$\frac{9031}{720}c^5$	$\frac{3211}{112}c^6$	$\frac{2685817}{40320}c^7$	$\frac{56953279}{362880}c^8$	$\frac{1357947691}{3628800}c^9$
$\frac{P_{\rm FC}(K=2,S\le s)}{\pi_{\rm FC}(K=2)}-1=$		$-2c$	$-4c^2$	$-\frac{25}{3}c^3$	$-18c^4$	$-\frac{2401}{60}c^5$	$-\frac{4096}{45}c^6$	$-\frac{59049}{280}c^7$	$-\frac{31250}{63}c^8$
$\frac{P_{\rm FC}(K=3,S\le s)}{\pi_{\rm FC}(K=3)}-1=$			$-3c$	$-\frac{15}{2}c^2$	$-18c^3$	$-\frac{343}{8}c^4$	$-\frac{512}{5}c^5$	$-\frac{19683}{80}c^6$	$-\frac{12500}{21}c^7$
$\frac{P_{\rm FC}(K=4,S\le s)}{\pi_{\rm FC}(K=4)}-1=$				$-4c$	$-12c^2$	$-\frac{98}{3}c^3$	$-\frac{256}{3}c^4$	$-\frac{2187}{10}c^5$	$-\frac{5000}{9}c^6$
$\frac{P_{\rm FC}(K=5,S\le s)}{\pi_{\rm FC}(K=5)}-1=$					$-5c$	$-\frac{35}{2}c^2$	$-\frac{160}{3}c^3$	$-\frac{1215}{8}c^4$	$-\frac{1250}{3}c^5$
$\frac{P_{\rm FC}(K=6,S\le s)}{\pi_{\rm FC}(K=6)}-1=$						$-6c$	$-24c^2$	$-81c^3$	$-250c^4$
$\frac{P_{\rm FC}(K=7,S\le s)}{\pi_{\rm FC}(K=7)}-1=$							$-7c$	$-\frac{63}{2}c^2$	$-\frac{350}{3}c^3$
$\frac{P_{\rm FC}(K=8,S\le s)}{\pi_{\rm FC}(K=8)}-1=$								$-8c$	$-40c^2$
$\frac{P_{\rm FC}(K=9,S\le s)}{\pi_{\rm FC}(K=9)}-1=$									$-9c$

TABLE V. The probabilities $P_{\rm FC}(L=\ell|L<\infty, S=s)$ that a pair of random nodes on a random cluster of size s in a subcritical ER network will be at a distance ℓ from each other for small tree cluster of $s=2,3,\dots,10$ nodes.

	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$	$s=9$	$s=10$
$P_{\rm FC}(L=1\|S=s)=$	1	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{3}$	$\frac{2}{7}$	$\frac{1}{4}$	$\frac{2}{9}$	$\frac{1}{5}$
$P_{\rm FC}(L=2\|S=s)=$		$\frac{1}{3}$	$\frac{3}{8}$	$\frac{9}{25}$	$\frac{1}{3}$	$\frac{15}{49}$	$\frac{9}{32}$	$\frac{7}{27}$	$\frac{6}{25}$
$P_{\rm FC}(L=3\|S=s)=$			$\frac{1}{8}$	$\frac{24}{125}$	$\frac{2}{9}$	$\frac{80}{343}$	$\frac{15}{64}$	$\frac{56}{243}$	$\frac{28}{125}$
$P_{\rm FC}(L=4\|S=s)=$				$\frac{6}{125}$	$\frac{5}{54}$	$\frac{300}{2401}$	$\frac{75}{512}$	$\frac{350}{2187}$	$\frac{21}{125}$
$P_{\rm FC}(L=5\|S=s)=$					$\frac{1}{54}$	$\frac{720}{16807}$	$\frac{135}{2048}$	$\frac{560}{6561}$	$\frac{63}{625}$
$P_{\rm FC}(L=6\|S=s)=$						$\frac{120}{16807}$	$\frac{315}{16384}$	$\frac{1960}{59049}$	$\frac{147}{3125}$
$P_{\rm FC}(L=7\|S=s)=$							$\frac{45}{16384}$	$\frac{4480}{531441}$	$\frac{252}{15625}$
$P_{\rm FC}(L=8\|S=s)=$								$\frac{560}{531441}$	$\frac{567}{156250}$
$P_{\rm FC}(L=9\|S=s)=$									$\frac{63}{156250}$
$\mathbb{E}[L\|S=s]=$	1	$\frac{4}{3}$	$\frac{13}{8}$	$\frac{236}{125}$	$\frac{115}{54}$	$\frac{39572}{16807}$	$\frac{42037}{16384}$	$\frac{1469756}{531441}$	$\frac{461843}{156250}$
$\mathbb{E}[L^2\|S=s]=$	1	2	$\frac{25}{8}$	$\frac{542}{125}$	$\frac{101}{18}$	$\frac{116582}{16807}$	$\frac{136033}{16384}$	$\frac{1718890}{177147}$	$\frac{1739471}{156250}$
$\mathrm{Var}[L\|S=s]=$	0	$\frac{2}{9}$	$\frac{31}{64}$	$\frac{12054}{15625}$	$\frac{3137}{2916}$	$\frac{393450490}{282475249}$	$\frac{461655303}{268435456}$	$\frac{580283161934}{282429536481}$	$\frac{58493387101}{24414062500}$

TABLE VI. The probabilities $P_{\mathrm{FC}}(L=\ell|L<\infty, S\leq s)$ that a pair of random nodes on a random cluster of size $S\leq s$ in a subcritical ER network will be at a distance ℓ from each other for small tree clusters of $s=2,3,\dots,10$ nodes.

	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$
$P_{\mathrm{FC}}(L=1\|S\leq s)=$	1	$\frac{1+2\eta}{1+3\eta}$	$\frac{1+2\eta+4\eta^2}{1+3\eta+8\eta^2}$	$\frac{6+12\eta+24\eta^2+50\eta^3}{6+18\eta+48\eta^2+125\eta^3}$	$\frac{6+12\eta+24\eta^2+50\eta^3+108\eta^4}{6+18\eta+48\eta^2+125\eta^3+324\eta^4}$
$P_{\mathrm{FC}}(L=2\|S\leq s)=$		$\frac{\eta}{1+3\eta}$	$\frac{\eta+3\eta^2}{1+3\eta+8\eta^2}$	$\frac{6\eta+18\eta^2+45\eta^3}{6+18\eta+48\eta^2+125\eta^3}$	$\frac{6\eta+18\eta^2+45\eta^3+108\eta^4}{6+18\eta+48\eta^2+125\eta^3+324\eta^4}$
$P_{\mathrm{FC}}(L=3\|S\leq s)=$			$\frac{\eta^2}{1+3\eta+8\eta^2}$	$\frac{6\eta^2+24\eta^3}{6+18\eta+48\eta^2+125\eta^3}$	$\frac{6\eta^2+24\eta^3+72\eta^4}{6+18\eta+48\eta^2+125\eta^3+324\eta^4}$
$P_{\mathrm{FC}}(L=4\|S\leq s)=$				$\frac{6\eta^3}{6+18\eta+48\eta^2+125\eta^3}$	$\frac{6\eta^3+30\eta^4}{6+18\eta+48\eta^2+125\eta^3+324\eta^4}$
$P_{\mathrm{FC}}(L=5\|S\leq s)=$					$\frac{6\eta^4}{6+18\eta+48\eta^2+125\eta^3+324\eta^4}$
$\mathbb{E}[L\|S\leq s]=$	1	$\frac{1+4\eta}{1+3\eta}$	$\frac{1+4\eta+13\eta^2}{1+3\eta+8\eta^2}$	$\frac{6+24\eta+78\eta^2+236\eta^3}{6+18\eta+48\eta^2+125\eta^3}$	$\frac{6+24\eta+78\eta^2+236\eta^3+690\eta^4}{6+18\eta+48\eta^2+125\eta^3+324\eta^4}$
$\mathbb{E}[L^2\|S\leq s]=$	1	$\frac{1+6\eta}{1+3\eta}$	$\frac{1+6\eta+25\eta^2}{1+3\eta+8\eta^2}$	$\frac{6+36\eta+150\eta^2+542\eta^3}{6+18\eta+48\eta^2+125\eta^3}$	$\frac{6+36\eta+150\eta^2+542\eta^3+1818\eta^4}{6+18\eta+48\eta^2+125\eta^3+324\eta^4}$

TABLE VII. The leading finite size correction terms $r_{s,\ell}c^{s-\ell}$ of Eq. (107) for the DSPL over all the tree topologies with up to s nodes. The distribution $P_{\mathrm{FC}}(L=\ell|L<\infty)=(1-c)c^{\ell-1}$, given by Eq. (108), is the DSPL over all pairs of nodes which reside on the same cluster in the entire subcritical network. As s is increased, the correction term decreases as c^{s-2} and $P_{\mathrm{FC}}(L=\ell|L<\infty, S\leq s)$ converges towards $P_{\mathrm{FC}}(L=\ell|L<\infty)$.

	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$	$s=9$	$s=10$
$\frac{P_{\mathrm{FC}}(L=1\|L<\infty,S\leq s)}{P_{\mathrm{FC}}(L=2\|L<\infty)}-1=$	1	$4c^2$	$\frac{25}{2}c^3$	$36c^4$	$\frac{2401}{24}c^5$	$\frac{4096}{15}c^6$	$\frac{59049}{80}c^7$	$\frac{125000}{63}c^8$	$\frac{214358881}{40320}c^9$
$\frac{P_{\mathrm{FC}}(L=2\|L<\infty,S\leq s)}{P_{\mathrm{FC}}(L=2\|L<\infty)}-1=$		$-3c$	$-\frac{15}{2}c^2$	$-18c^3$	$-\frac{343}{8}c^4$	$-\frac{512}{5}c^5$	$-\frac{19683}{80}c^6$	$-\frac{12500}{21}c^7$	$-\frac{19487171}{13440}c^8$
$\frac{P_{\mathrm{FC}}(L=3\|L<\infty,S\leq s)}{P_{\mathrm{FC}}(L=3\|L<\infty)}-1=$			$-4c$	$-12c^2$	$-\frac{98}{3}c^3$	$-\frac{256}{3}c^4$	$-\frac{2187}{10}c^5$	$-\frac{5000}{9}c^6$	$-\frac{1771561}{1260}c^7$
$\frac{P_{\mathrm{FC}}(L=4\|L<\infty,S\leq s)}{P_{\mathrm{FC}}(L=4\|L<\infty)}-1=$				$-5c$	$-\frac{35}{2}c^2$	$-\frac{160}{3}c^3$	$-\frac{1215}{8}c^4$	$-\frac{1250}{3}c^5$	$-\frac{161051}{144}c^6$
$\frac{P_{\mathrm{FC}}(L=5\|L<\infty,S\leq s)}{P_{\mathrm{FC}}(L=5\|L<\infty)}-1=$					$-6c$	$-24c^2$	$-81c^3$	$-250c^4$	$-\frac{14641}{20}c^5$
$\frac{P_{\mathrm{FC}}(L=6\|L<\infty,S\leq s)}{P_{\mathrm{FC}}(L=6\|L<\infty)}-1=$						$-7c$	$-\frac{63}{2}c^2$	$-\frac{350}{3}c^3$	$-\frac{9317}{24}c^4$
$\frac{P_{\mathrm{FC}}(L=7\|L<\infty,S\leq s)}{P_{\mathrm{FC}}(L=7\|L<\infty)}-1=$							$-8c$	$-40c^2$	$-\frac{484}{3}c^3$
$\frac{P_{\mathrm{FC}}(L=8\|L<\infty,S\leq s)}{P_{\mathrm{FC}}(L=8\|L<\infty)}-1=$								$-9c$	$-\frac{99}{2}c^2$
$\frac{P_{\mathrm{FC}}(L=9\|L<\infty,S\leq s)}{P_{\mathrm{FC}}(L=9\|L<\infty)}-1=$									$-10c$

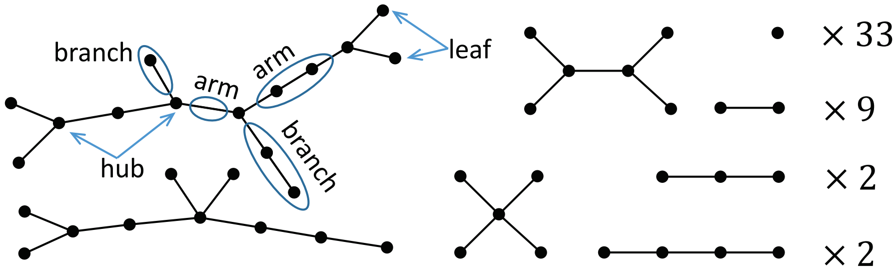

FIG. 1. (Color online) Illustration of the structure of one instance of a subcritical ER network of $N = 100$ nodes and $c = 0.9$. It consists of 33 isolated nodes, 9 dimers, two chains of three nodes, two chains of four nodes, one tree with a single hub and four branches, one tree with two hubs and two larger trees of 10 and 14 nodes.

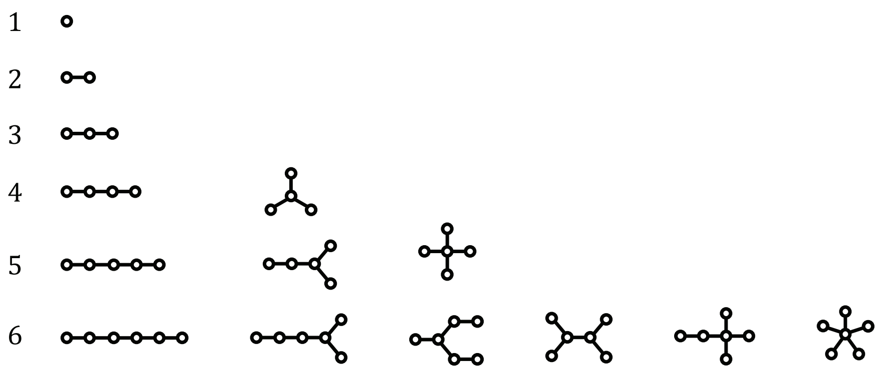

FIG. 2. A list of all the possible backbone tree topologies which consist of up to six hubs ($h = 1, 2, \ldots, 6$). The linear chain topology appears for all values of h. For $h \leq 3$ it is the only topology, while for $h \geq 4$ more complex topologies appear and their number quickly increases.

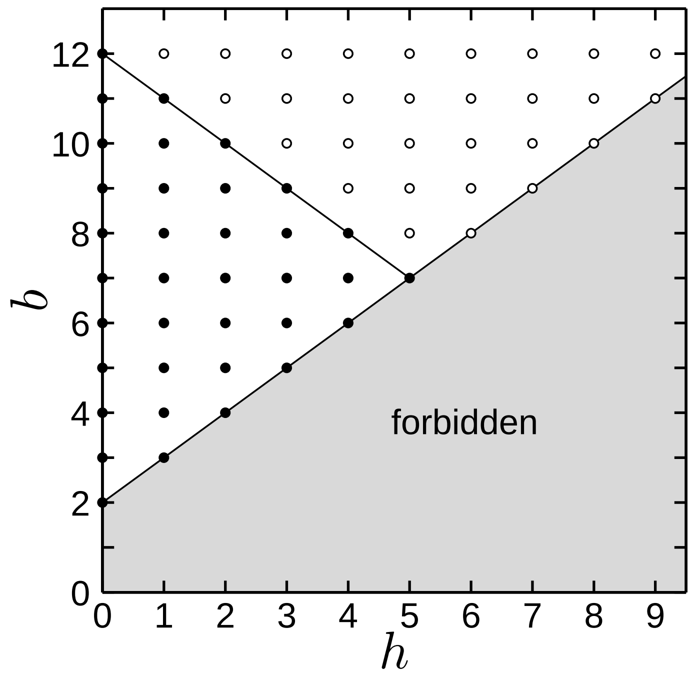

FIG. 3. Illustration of the range of possible values of b (the number of branches) in a tree of h hubs, which consists of s nodes. This range is bounded from below by $b = h + 2$, due to a topological constraint, regardless of s (ascending straight line). For a network which consists of s nodes, is is bounded from above by $b = s - h$ (descending straight line). The two lines intersect at $(h, 2h + 2)$. Combinations of (h, b) which are possible in a tree of $s = 12$ nodes are marked by full circles, while combination which exist only in larger trees are marked by empty circles. Each backbone tree can be represented by its adjacency matrix A, which is an $h \times h$ matrix. The topology of a complete tree is denoted by $\tau = (h, A, \vec{b})$, where $\vec{b} = (b_1, \ldots, b_h)$ accounts for a specific division of the b branches between the h hubs.

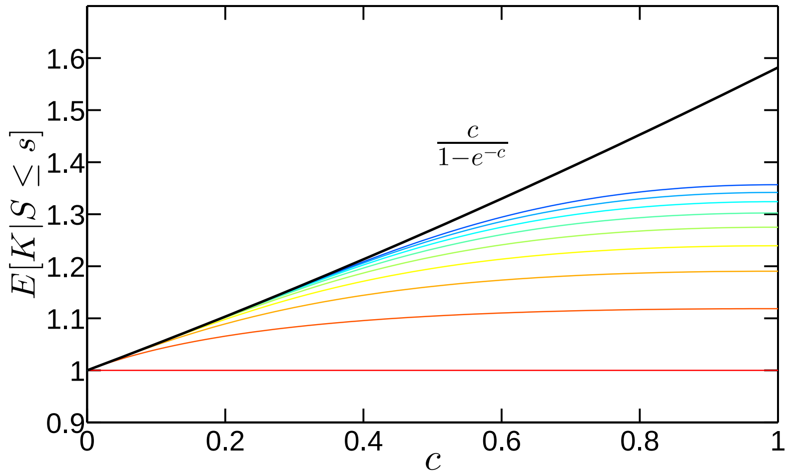

FIG. 4. (Color online) The mean degree, $\mathbb{E}[K|S \le s]$, over all tree topologies of sizes smaller or equal to s, as a function of c, for $s = 2, 3, \ldots, 10$ (solid lines), from bottom to top, respectively. The thick solid line shows the asymptotic result, $\langle K \rangle_{\rm FC}$.

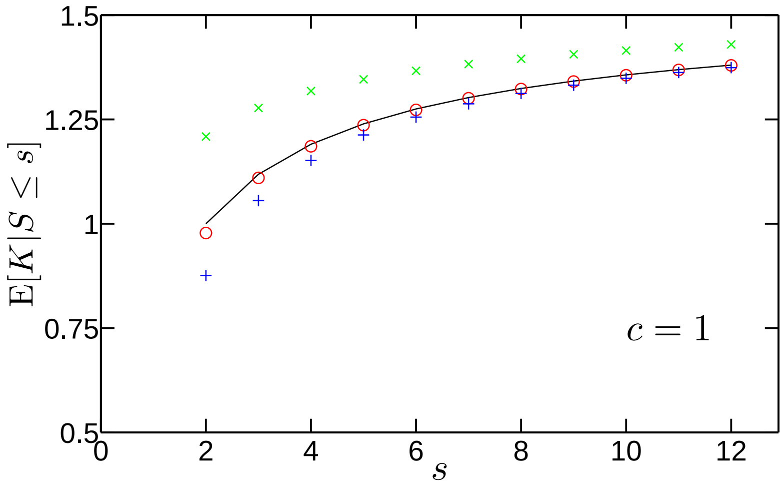

FIG. 5. (Color online) The mean degree $\mathbb{E}[K|S \le s]$ over all trees of size smaller or equal to s, as a function of s for $c = 1$. The analytical results (circles), obtained from Eq. (85), are in excellent agreement with the exact results of the asymptotic expansion (solid line). The results of the asymptotic expansion to order $1/\sqrt{s}$ ($\times$), obtained from the first two terms of Eq. (86), exhibit large deviations from the exact results, particularly for small values of s. However, an expansion to order $1/s$ obtained by including the third term in Eq. (86), greatly improves the results (+).

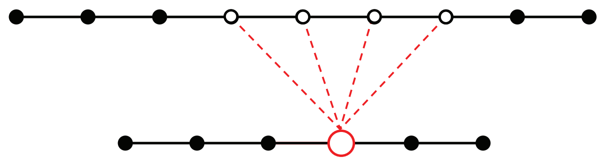

FIG. 6. (Color online) Illustration of the collapse process which is used in order to obtain the combinatorial factors for the DSPL on a finite cluster. In this case, the number of pairs of nodes which are at a distance of $\ell = 3$ from each other on a linear chain of size $s = 9$ (top chain) is equal to the number of possible locations of the marked node (large empty circle) on the reduced chain of $s - \ell = 6$ nodes (bottom chain).

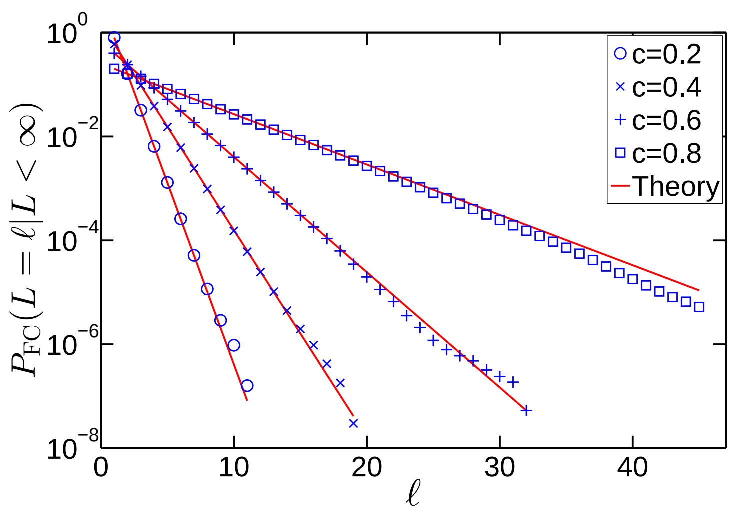

FIG. 7. (Color online) The DSPLs of subcritical ER network ensembles with $N = 10^4$ and $c = 0.2, 0.4, 0.6$ and 0.8. The theoretical results for the corresponding asymptotic networks (solid lines), obtained from Eq. (108) are in very good agreement with the numerical simulations (symbols). The deviations in the tail are due to the finite size of the sampled networks.

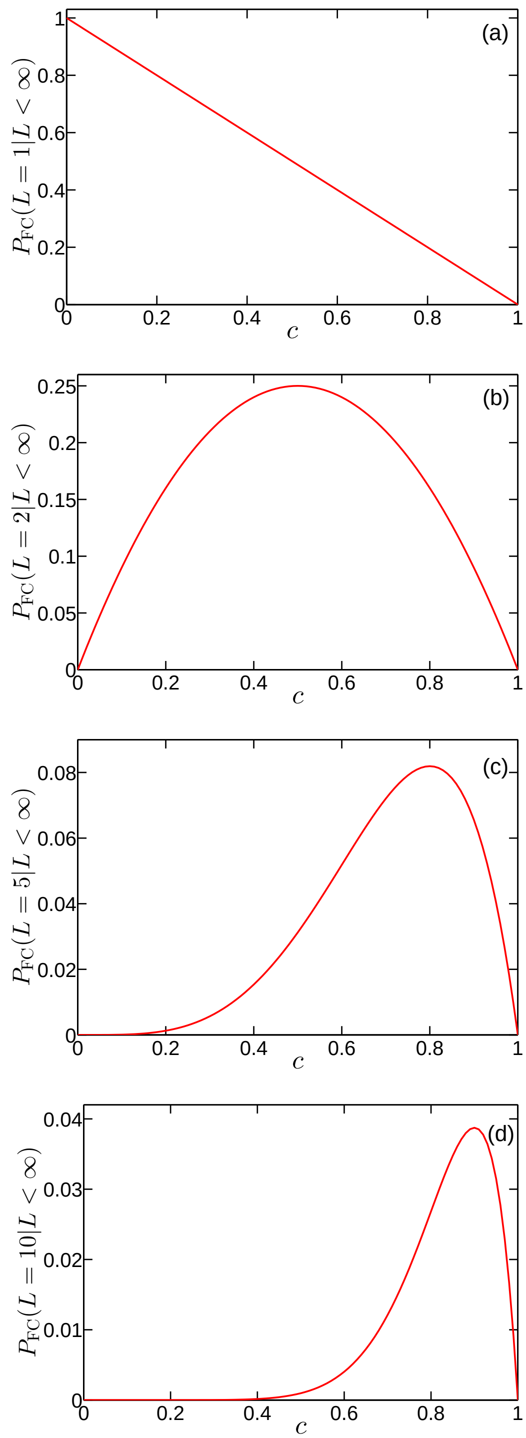

FIG. 8. (Color online) The probability $P_{\rm FC}(L=\ell|L<\infty)$ given by Eq. (108), is shown as a function of the mean degree, c, for $\ell=1,2,5$ and 10. The probability $P_{\rm FC}(L=1|L<\infty)$ is a monotonically decreasing function of c. For $\ell\geq 2$, the probability $P_{\rm FC}(L=\ell|L<\infty)$ exhibits a peak at $c=1-1/\ell$, which is the value of c at which the probability of two random nodes to be at a distance ℓ from each other is maximal.

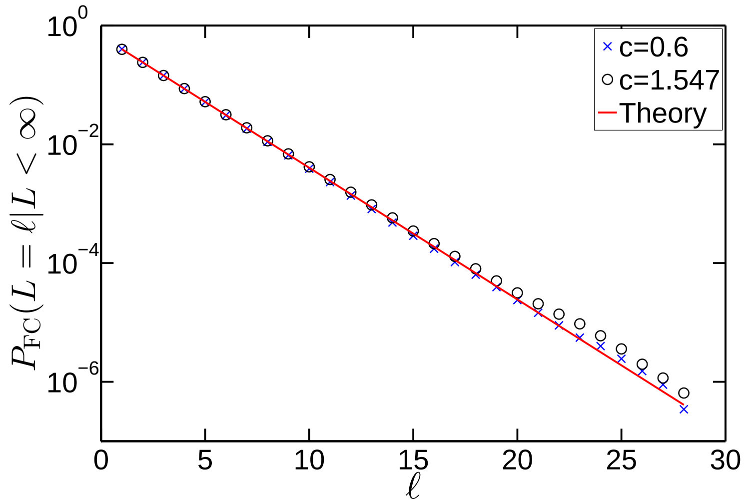

FIG. 9. (Color online) Numerical results for the DSPL on the finite clusters of an ER$[N, c/(N-1)]$ network with $N = 10^4$ and $c = 1.547$, above percolation (circles) and on its dual network, ER$[N', c'/(N'-1)]$ where $N' = 3882$ [obtained from Eq. (117)] and $c' = 0.6$ [obtained from Eq. (118)], below percolation ($\times$), which are essentially identical and in excellent agreement with the theoretical results (solid line), obtained from Eq. (108).

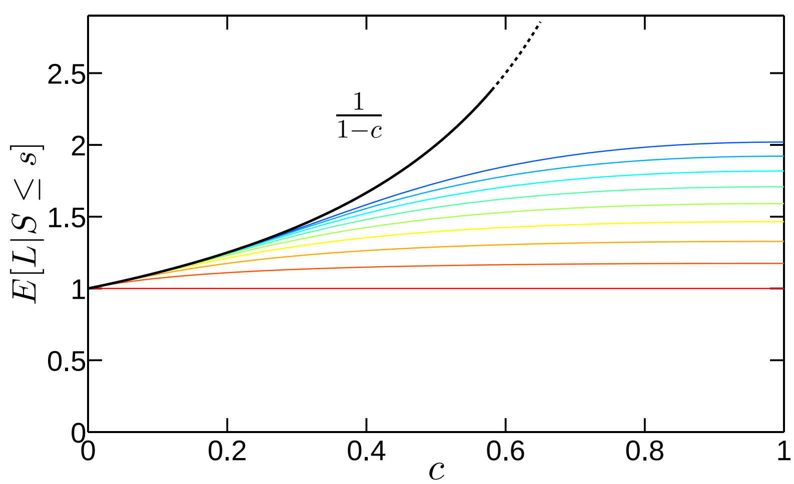

FIG. 10. (Color online) The mean distance $\mathbb{E}[L|S \le s]$, over all tree topologies of sizes smaller or equal to s, as a function of c, for $s = 2, 3, \ldots, 10$ (solid lines), from bottom to top, respectively. The thick solid line shows the asymptotic result, $\langle L \rangle_{\rm FC}$.

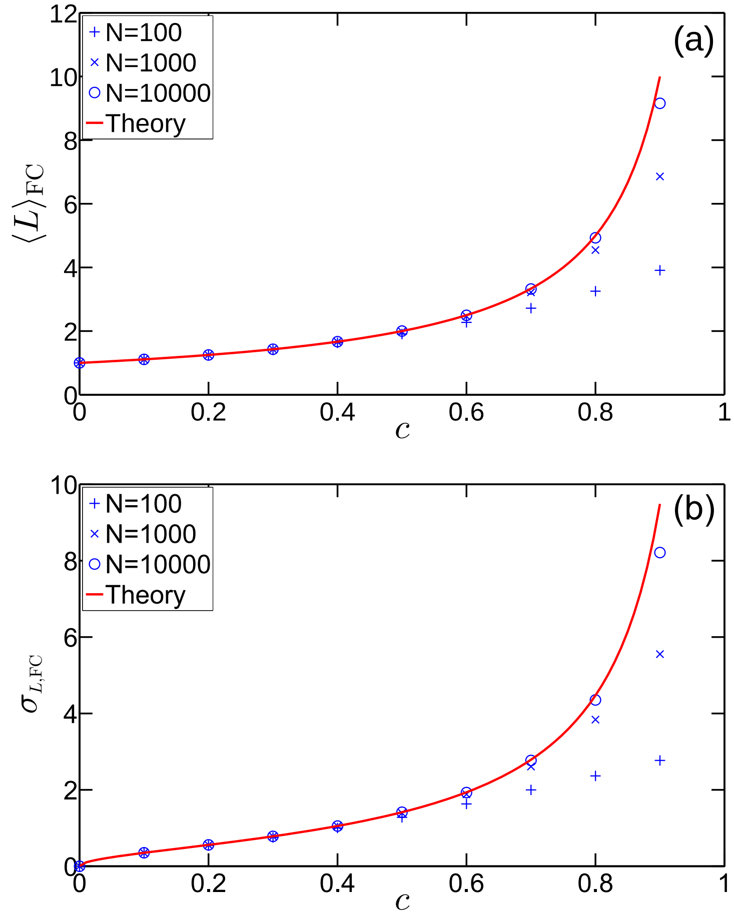

FIG. 11. (Color online) The mean, $\langle L \rangle_{\rm FC}$ (a), and the standard deviation, $\sigma_{L,\rm FC}$ (b), of the DSPL of a subcritical ER network vs. the mean degree, c. The numerical results (symbols) for $N = 10^2$, 10^3 and 10^4 clearly converge towards the analytical results (solid lines).