A Letters Journal Exploring the Frontiers of Physics

July 2015

EPL, **111** (2015) 26006
doi: 10.1209/0295-5075/111/26006

www.epljournal.org

Analytical results for the distribution of shortest path lengths in random networks

Eytan Katzav[1], Mor Nitzan[1,2], Daniel ben-Avraham[3], P. L. Krapivsky[4], Reimer Kühn[5], Nathan Ross[6] and Ofer Biham[1]

[1] *Racah Institute of Physics, The Hebrew University - Jerusalem 91904, Israel*
[2] *Department of Microbiology and Molecular Genetics, Faculty of Medicine, The Hebrew University Jerusalem 91120, Israel*
[3] *Department of Physics, Clarkson University - Potsdam, NY 13699-5820, USA*
[4] *Department of Physics, Boston University - Boston, MA 02215, USA*
[5] *Department of Mathematics, King's College London - Strand, London WC2R 2LS, UK*
[6] *School of Mathematics and Statistics, University of Melbourne - Melbourne, VIC 3010, Australia*

received 21 April 2015; accepted 20 July 2015
published online 13 August 2015

PACS 64.60.aq – Networks
PACS 89.75.Da – Systems obeying scaling laws

Abstract – We present two complementary analytical approaches for calculating the distribution of shortest path lengths in Erdős-Rényi networks, based on recursion equations for the shells around a reference node and for the paths originating from it. The results are in agreement with numerical simulations for a broad range of network sizes and connectivities. The average and standard deviation of the distribution are also obtained. In the case in which the mean degree scales as N^{α} with the network size, the distribution becomes extremely narrow in the asymptotic limit, namely almost all pairs of nodes are equidistant, at distance $d = \lfloor 1/\alpha \rfloor$ from each other. The distribution of shortest path lengths between nodes of degree m and the rest of the network is calculated. Its average is shown to be a monotonically decreasing function of m, providing an interesting relation between a local property and a global property of the network. The methodology presented here can be applied to more general classes of networks.

The increasing interest in network research in recent years is motivated by the realization that a large variety of systems and processes which involve interacting objects can be described by network models [1–4]. In these models, the objects are represented by nodes and the interactions are expressed by edges. Pairs of connected nodes can affect each other directly. However, the interactions between most pairs of nodes are indirect, mediated by intermediate nodes and edges. Important properties of these indirect interactions such as their strengths, delay times, coordination, correlation and synchronization depend on the paths between different nodes. A pair of nodes, i and j, may be connected by a large number of paths. The shortest among these paths are of particular importance because they are likely to provide the fastest and strongest interaction between these two nodes. Therefore, it is of interest to study the distribution of shortest path lengths (DSPL) between nodes in different types of networks. Such distributions are expected to depend on the network structure and size.

Random networks of the Erdős-Rényi (ER) type were studied extensively since the 1950s [5–7] using mathematical methods and computer simulations [8]. The increasing availability of empirical data on networks since the late 1990s stimulated much theoretical interest, leading to new results for ER networks [9,10]. Measures such as the diameter and the average path length were studied extensively [11,12]. However, apart from a few studies, the entire DSPL has attracted little attention [13–15]. This distribution is of great importance for the temporal evolution of dynamical processes on networks, such as signal propagation, navigation and epidemic spreading [16]. It determines the number of nodes exposed to a propagating signal originated from a given node as a function of time. More generally, the shortest paths can be considered as the backbone of a more complete set of paths between pairs of

nodes. While the shortest paths provide the fastest propagation, signals also utilize longer paths which are more numerous. This was demonstrated in studies of first passage times in diffusive processes on networks [17].

In this letter we present two analytical approaches for calculating the DSPL between nodes in the ER network, referred to as the recursive shells approach (RSA) and the recursive paths approach (RPA). Using recursion equations we study this distribution in different regimes, namely sparse and dense networks of small as well as asymptotically large sizes. Consider an ER network of N nodes, where each pair of nodes is independently connected with probability p. We denote such a network by ER(N,p). Its degree sequence follows the Poisson distribution with the parameter Np, which is equal to the average degree. Such networks are often studied in the asymptotic limit, where $N \to \infty$. In this limit, one can identify different regimes, according to the scaling of p *vs.* N.

For sparse networks, denoted by ER$(N, c/N)$, the average degree is $c = Np$. At $c = 1$ there is a percolation transition. For $c < 1$, the network consists of small isolated clusters. For $c > 1$, a giant component of size which scales linearly with N is formed, in addition to the small, isolated components of maximal size which scales as $\ln N$ [8]. For dense networks, the parameter p scales as $N^{\alpha-1}$, where $0 < \alpha < 1$, the mean connectivity grows with the network size as N^{α} and the number of isolated components vanishes.

When a pair of nodes resides on the same connected subnetwork, one can identify paths connecting these nodes. The path length is the number of edges along the path. The distance d_{ij} between a pair of different nodes i and j is the length of the shortest path connecting them. When i and j reside on different sub-networks, there is no path between them and thus $d_{ij} \equiv \infty$. The tail distribution $F_N(k) = Pr(d > k)$, $k = 0, 1, 2, \ldots, N-1$, is the probability that the distance d between a random pair of nodes in an ER network of size N is larger than k. Clearly, the probability that two distinct random nodes are at a distance $d > 0$ from each other is $F_N(0) = 1$, while the probability that $d > 1$, *i.e.* they are not directly connected, is $F_N(1) = q$, where $q = 1 - p$. The probability distribution $P_N(k)$ can be recovered as $P_N(k) = F_N(k-1) - F_N(k)$, $k = 1, 2, \ldots, N-1$. The probability $F_N(k)$ does not necessarily converge to zero in the limit $k \to \infty$. Its asymptotic value $F(\infty)$ is equal to the fraction of pairs of nodes in the network which belong to different clusters, namely for which $d_{ij} = \infty$. In fact, $F(\infty)$ can be estimated independently by using known properties of the fraction of nodes, g, which belongs to the giant component in the asymptotic limit [8]. This fraction satisfies $g = 1 - \exp(-cg)$ and $F(\infty) = 1 - g^2$. In a finite network $F(\infty)$ can be replaced by $F_N(N-1)$ since the longest possible distance is $d = N - 1$.

In the RSA, one picks a random node, i, as a reference node and examines the shell structure of the rest of the network around it. The number of nodes which are at a distance $d > k$, $k = 0, 1, 2, \ldots, N-1$, from the reference node is denoted by $\overline{N}_k$. The number of nodes at distance $d = k$ from the reference node is denoted by N_k, where $N_0 = 1$ and $N_k = \overline{N}_{k-1} - \overline{N}_k$ for $k \geq 1$. The N_k's obey the recursion equation $N_{k+1} = \overline{N}_k(1 - q^{N_k})$, which can be re-written as a second-order difference equation of the form $\overline{N}_{k+1} = \overline{N}_k q^{\overline{N}_{k-1} - \overline{N}_k}$, where $\overline{N}_0 = N - 1$ and $\overline{N}_1 = (N-1)q$. Using the relation $\overline{N}_k = (N-1) \cdot F_N(k)$, it can be expressed as

$$F_N(k+1) = F_N(k) q^{(N-1)[F_N(k-1) - F_N(k)]}, \qquad (1)$$

where $F_N(0) = 1$ and $F_N(1) = q$.

In the RPA one first picks two distinct random nodes, i and j. The probability that the distance between them is larger than k can be related to the probability that it is larger than $k-1$ by $F_N(k) = F_N(k-1) P_N(d > k | d > k-1)$, where $P_N(d > k | d > k-1)$ is the conditional probability that the distance is larger than k, given that it is larger than $k-1$. The iteration of this relation yields

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

This means that in order to obtain the distribution $F_N(k)$, all we need to calculate are the conditional probabilities $P_N(d > m | d > m-1)$, for all values of $2 \leq m \leq k$.

Consider a path of length k starting at node i and ending at node j (assuming that there is no such path of length $k-1$ or less). The path can be decomposed into a single edge from node i to an intermediate node ℓ and a shorter path of length $k-1$ from ℓ to j. Such a path can be ruled out in two ways: either there is no edge between i and ℓ (with probability q), or, in case that there is such an edge, there is no path of length $k-1$ between ℓ and j. The probability of the latter is $P_{N-1}(d > k-1 | d > k-2)$, since the remaining path is embedded in a smaller network of $N-1$ nodes. Combining the two possibilities yields the recursion equation

$$P_N(d > k | d > k-1) = \left[q + p \cdot P_{N-1}(d > k-1 | d > k-2) \right]^{N-2}, \qquad (3)$$

where the right-hand side is raised to the power $N-2$ in order to account for all possible ways to choose the intermediate node ℓ. In fig. 1 we present the possible paths of length k between i and j. This approach follows the spirit of the renormalization group theory [18], since the removal of a node from the network reduces the size of the configuration space by a factor of 2^{N-1}. This process is repeated $k-1$ times, reducing the network down to size $N' = N - k + 1$ and closing the recursion equations with $P_{N'}(d > 1 | d > 0) = F_{N'}(1) = q$.

Interestingly, inserting $k = 2$ in eq. (3) gives rise to the simple and exact expression

$$P_N(d > 2 | d > 1) = (1 - p^2)^{N-2}. \qquad (4)$$

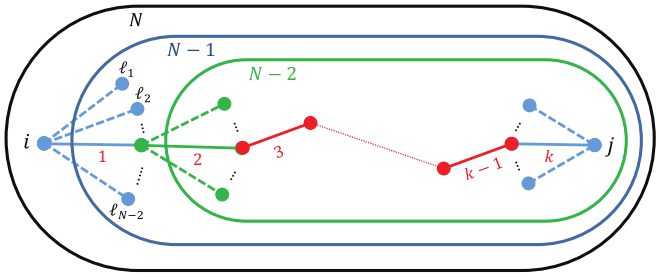

Fig. 1: (Color online) Illustration of the possible paths of length k between two random nodes i and j in an ER network of N nodes. The first edge of such path connects node i to some other node ℓ, which may be any one of the remaining $N-2$ nodes. The rest of the path, from ℓ to j is of length $k-1$ and it resides on a smaller network of $N-1$ nodes. The path of length k from i to j exists only when both the edge from i to ℓ and the path of length $k-1$ from ℓ to j exist.

Each path of length $k=2$ between nodes i and j consists of a single intermediate node and two edges. These paths do not overlap and are thus independent. Paths of lengths $k>2$ may share edges with other paths of the same length as well as with shorter paths. Therefore, in the calculation of the DSPL we use conditional probabilities to ensure that no shorter paths exist. This approach eliminates the correlations between paths of different lengths. On the other hand, nodes i and j may be connected by several paths of the same length, which may share some edges and thus become correlated. The RPA does not account for such correlations, because it assumes that the sub-networks of size $N-1$ are independent. Averaging over the quenched randomness in each instance of such network, the RPA provides the distribution over an ensemble of networks.

In the limit $p \to 0$ one can simplify the recursion equations and obtain the approximate closed form expression

$$P_N(d>k|d>k-1) = (1-p^k)^{(N-2)\dots(N-k)}, \qquad (5)$$

for any value of k. This expression is obtained using induction, based on eq. (3) and the exact result given above for $k=2$. This can be understood intuitively since the total number of possible paths of length k between nodes i and j is given by the product $(N-2)\dots(N-k)$, and the probability for each of these paths to be connected is given by p^k. This approximation breaks down for values of p which are not exceedingly small, where the correlations between different paths build up and cannot be ignored.

The regime of sparse networks was studied extensively, focusing on the diameter (namely, the largest distance between any pair of nodes) of the giant cluster, which scales like a constant times $\ln N$, where the constant is $1/\ln c - 2/\ln c'$, where $c' < 1$ satisfies the equation $c'\exp(-c') = c\exp(-c)$ [19]. In the strongly connected regime, we focus on the case in which $p = bN^{\alpha-1}$, where $b>0$ and $0<\alpha<1$. In this case the average degree increases with the network size as N^α. We will now derive an asymptotic result for the limit $N\to\infty$. In this limit $p\to 0$ and therefore the simplified results of eq. (5) can be used. Plugging the scaling of p *vs.* N into eq. (5) one obtains

$$P_N(d>k|d>k-1) \simeq \left(1-\frac{b^k}{N^{k(1-\alpha)}}\right)^{N^{k-1}}. \qquad (6)$$

For $N\to\infty$, $P_N(d>k|d>k-1) \to P(d>k|d>k-1)$, where $P(d>k|d>k-1)=1$ for $k<1/\alpha$, $\exp(-b^{1/\alpha})$ for $k=1/\alpha$ and 0 for $k>1/\alpha$. Note that the second case in the above equation is obtained only in the special case of $\alpha=1/r$, where r is an integer. Therefore, we will first consider the generic case in which α is not an exact inverse of an integer. Inserting the result for the conditional probabilities into eq. (2) we obtain

$$P(k) = \begin{cases} 1, & k = \lfloor \frac{1}{\alpha} \rfloor + 1, \\ 0, & \text{otherwise}, \end{cases} \qquad (7)$$

where $\lfloor x \rfloor$ is the integer part of x. In case that $\alpha = 1/r$ we obtain that

$$P(k) = \begin{cases} 1-e^{-b^r}, & k=r, \\ e^{-b^r}, & k=r+1, \\ 0, & \text{otherwise}. \end{cases} \qquad (8)$$

These results can be understood intuitively using the following argument. Starting from node i, we define the shell of radius $d=1$ around it as the set of nodes which are directly connected to i. The expected value for the number of nodes in this shell is $N_1 \sim N^\alpha$. Proceeding by induction, the shell of radius d is denoted as the set of nodes which are directly connected to nodes in the shell of radius $d-1$. Thus, the number of nodes in the shell of radius d is given by $N_d \sim N^{d\alpha}$. In the asymptotic limit, as long as $d\alpha<1$, the shell of radius d still consists of an exceedingly small fraction of the nodes in the network. On the other hand, once $d\alpha>1$, this shell includes almost every node in the network. This means that almost all the nodes in the network are at a distance $d=\lfloor 1/\alpha \rfloor + 1$ from node i. Since node i was chosen at random, this means that the shortest path between almost any pair of nodes in the network is of length d.

The case of $\alpha = 1/r$, where r is an integer, requires a special consideration. Based on the argument presented above, the neighborhood of radius $d=r$ from node i should include all the N nodes. However, this counting includes duplications, namely nodes which are connected to node i by several paths of length r. As a result, there are other nodes which are not reached by any of these paths. Since the number of nodes of distance r from node i scales with N, it is clear that each one of the remaining nodes is connected to at least one of them. Therefore, the remaining nodes are at a distance $d=r+1$ from node i.

Before presenting the results obtained from the two approaches, we refer to an earlier study of the DSPL

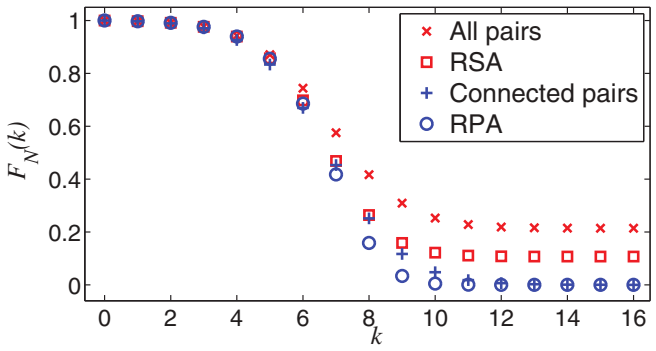

Fig. 2: (Color online) (a) The tail distribution $F_N(k)$ *vs.* k for the ER$(N, c/N)$ network with $N = 1000$ and $c = 2.5$, obtained from numerical simulations for all pairs of nodes ($\times$) and for pairs of nodes on the same cluster ($+$). The results of the RSA ($\square$) agree well with the numerical results for all pairs of nodes, except for the asymptotic tail. The results of the RPA ($\circ$) agree well with the numerical results for pairs of nodes on the same cluster.

in ER networks [13]. We briefly summarize their approach, adapting the notation where appropriate. The expectation value for the number of nodes at a distance $k-1$ or less from the reference node is given by $n(k) = [1 - F_N(k-1)]N$. This is due to the fact that the probability for a random node to be at a distance smaller than k is $(1 - F_N(k-1))$, and multiplying by N one obtains $n(k)$. In order for a node to be at a distance larger than k from the reference node, it must not be directly connected to any of the $n(k)$ nodes which are at distance $k-1$ or less from the reference node. Picking a random node, the probability that it will not be connected to any of these nodes is given by [13]

$$F_N(k) = q^{[1-F_N(k-1)]N}. \tag{9}$$

This recursion equation can be iterated, starting from $F_N(0) = (N-1)/N$, to obtain $F_N(k)$ for $k = 1, 2, \ldots$. A potential problem with this approach is that in the estimation of the probability, $F_N(k)$, that a random node will be at distance larger than k from the reference node, eq. (9) ignores the possibility that the random node is already connected to the reference node by a path of length $k-1$ or less. This is expected to bias the distribution towards larger distances.

In fig. 2 we present the tail distribution $F_N(k)$ *vs.* k, for an ER network of $N = 1000$ nodes and $p = c/N$, where $c = 2.5$, obtained from numerical simulations for all pairs of nodes ($\times$) and for pairs of nodes on the same cluster ($+$). We also present the theoretical results obtained from the RSA ($\square$) and from the RPA ($\circ$). The results of the RSA agree well with the numerical results for all pairs, except for the limit of large distances where the plateau in $F_N(k)$ is lower than the empirical curve. This means that this approach underestimates the fraction of pairs for which $d_{ij} = \infty$, which is equal to $F(\infty)$. The results of the RPA agree well with the numerical results for pairs which reside on the same cluster. This is due to the fact

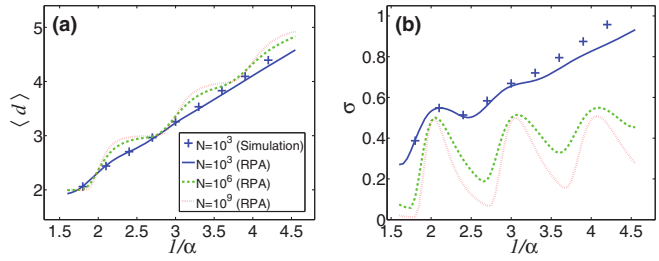

Fig. 3: (Color online) The average $\langle d \rangle$ (a) and the standard deviation σ (b) of the DSPL in the ER$(N, bN^{\alpha-1})$ network *vs.* $1/\alpha$ for $b = 1$ and $N = 10^3$ (solid line) 10^6 (dashed line) and 10^9 (dotted line), obtained from the RPA. It is observed that $\langle d \rangle \simeq \lfloor 1/\alpha \rfloor + 1$, decorated by a rounded step function, while σ exhibits oscillations with maxima at integer values of $1/\alpha$.

that this approach reconstructs the remaining network at each iteration of the recursion equations. As a result, the quenched randomness of the connectivities in each realization of the network is annealed, eliminating the isolated nodes and the small, isolated clusters. In the RSA there is no such annealing. Therefore, the RSA applies to all pairs of nodes in the network while the RPA applies to pairs of nodes on the same cluster. In the limit of dense networks there are no isolated components and the two approaches coincide.

The distribution $P_N(k)$ can be characterized by its moments. The n-th moment, $\langle k^n \rangle$, can be obtained using the tail-sum formula $\langle k^n \rangle = \sum_{k=0}^{N-1}[(k+1)^n - k^n]F_N(k)$. In particular, the first moment is given by $\langle k \rangle = \sum_{k=0}^{N-1} F_N(k)$ and the second moment by $\langle k^2 \rangle = \sum_{k=0}^{N-1}(2k+1)F_N(k)$. The width of the distribution can be characterized by the variance $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$. Related topological indices [20] such as the Wiener index [21] and the Harary index [22–24] were studied in the context of chemical graphs. It was shown that important properties of molecules can be obtained using such indices for the graphs representing their structure [21].

In fig. 3(a) we present the average distance $\langle d \rangle$ between pairs of nodes *vs.* $1/\alpha$ in dense ER networks. Following eqs. (7), (8), these functions converge to a staircase form as $N \to \infty$. In fig. 3(b) we present the standard deviation σ *vs.* $1/\alpha$. For finite networks it exhibits oscillations of unit period. In the asymptotic limit the peaks become vanishingly narrow around the integers.

So far we have studied the DSPL between all pairs of nodes in the network. Below, we consider a reference node i of a known degree, m, and study the DSPL between this node and the rest of the network. We denote the DSPL between a random node i of degree m and other random nodes, j, by $F_{N|m}(k) = F_N(k|\deg(i) = m)$ and the corresponding conditional probability by $P_{N|m}(d > k|d > k-1)$. In this case, the first iteration of the recursion equation takes the form

$$P_{N|m}(d > k|d > k-1) = [P_{N-1}(d > k-1|d > k-2)]^m, \tag{10}$$

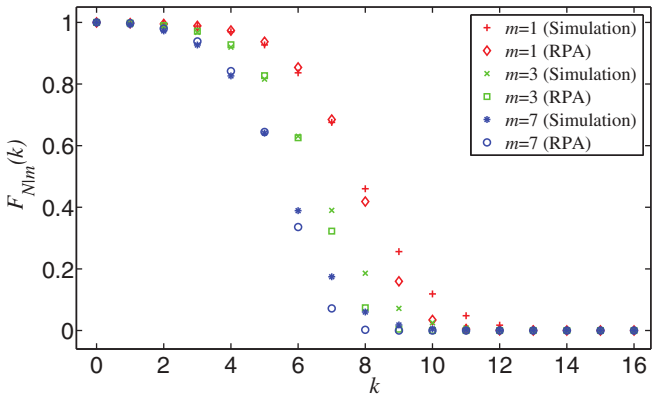

Fig. 4: (Color online) The DSPL $F_{N|m}(k)$ *vs.* k between a random node i of a given degree, m, and all other nodes which reside on the same cluster in a dilute ER network of $N = 1000$ and $c = 2.5$. The results of the RPA for $m = 1$ ($\diamond$), 3 ($\square$) and 7 ($\circ$) are in good agreement with the corresponding numerical results: $m = 1$ ($+$), 3 ($\times$) and 7 ($*$).

where the expression on the right-hand side is obtained from eq. (3). In fig. 4 we present the tail distribution $F_{N|m}(k)$ *vs.* k, obtained from numerical simulations for $m = 1$ ($+$), 3 ($\times$) and 7 ($*$), in a dilute ER network of $N = 1000$ and $c = 2.5$. Each data point is averaged over 20 independent realizations of the network. The results of the RPA for $m = 1$ ($\diamond$), 3 ($\square$) and 7 ($\circ$) are in good agreement with the numerical results. Clearly, the distribution is strongly affected by the local connectivity of the reference node. The knee of the distribution $F_{N|m}(k)$ (which coincides with the peak of the corresponding probability density function) moves to the left as m is increased. This means that nodes which are strongly connected at the local level are closer to the rest of the network than weakly connected nodes.

In summary, we have studied the distribution of shortest path lengths in ER networks using two complementary theoretical approaches and showed that they are in good agreement with numerical results. For large and dense networks the distribution becomes extremely narrow and is exactly captured by both approaches. A slight modification enables us to calculate the DSPL around a node with a given degree, m. The results exemplify the impact of local features (such as the degree of a node) on global properties (such as the distance distribution) in complex networks. The proposed theoretical approaches are highly flexible and can be applied to more general networks [25,26].

* * *

DBA, OB and PLK acknowledge the Galileo Galilei Institute for Theoretical Physics in Florence and INFN for the hospitality and support during the workshop on Advances in Nonequilibrium Statistical Mechanics in Spring 2014, where preliminary work was performed. MN is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship.

REFERENCES

[1] ALBERT R. and BARABÁSI A. L., *Rev. Mod. Phys.*, **74** (2002) 47.
[2] CALDARELLI G., *Scale Free Networks: Complex Webs in Nature and Technology* (Oxford University Press, Oxford) 2007.
[3] NEWMAN M. E. J., *Networks: an Introduction* (Oxford University Press, Oxford) 2010.
[4] ESTRADA E., *The Structure of Complex Networks: Theory and Applications* (Oxford University Press, Oxford) 2011.
[5] ERDŐS P. and RÉNYI A., *Publ. Math.*, **6** (1959) 290.
[6] ERDŐS P. and RÉNYI A., *Publ. Math. Inst. Hung. Acad. Sci.*, **5** (1960) 17.
[7] ERDŐS P. and RÉNYI A., *Bull. Inst. Int. Stat.*, **38** (1961) 343.
[8] BOLLOBÁS B., *Random Graphs* (Cambridge University Press, Cambridge) 2001.
[9] VAN DER HOFSTAD R., HOOGHIEMSTRA G. and VAN MIEGHEM P., *Random Struct. Algorithms*, **27** (2005) 76.
[10] DE BACCO C., MAJUMDAR S. N. and SOLLICH P., *J. Phys. A: Math. Theor.*, **48** (2015) 205004.
[11] WATTS D. J. and STROGATZ S. H., *Nature*, **393** (1998) 440.
[12] FRONCZAK A., FRONCZAK P. and HOLYST J. A., *Phys. Rev. E*, **70** (2004) 056110.
[13] BLONDEL V. D., GUILLAUME J.-L., HENDRICKX J. M. and JUNGERS R. M., *Phys. Rev. E*, **76** (2007) 066101.
[14] DOROGOTSEV S. N., MENDES J. F. F. and SAMUKHIN A. N., *Nucl. Phys. B*, **653** (2003) 307.
[15] VAN DER ESKER H., VAN DER HOFSTAD R. and HOOGHIEMSTRA G., *J. Stat. Phys.*, **133** (2008) 169.
[16] PASTOR-SATORRAS R. and VESPIGNANI A., *Phys. Rev. Lett.*, **86** (2001) 3200.
[17] SOOD V., REDNER S. and BEN-AVRAHAM D., *J. Phys. A*, **38** (2005) 109.
[18] BINNEY J. J., DOWRICK N. J., FISHER A. J. and NEWMAN M. E. J., *The Theory of Critical Phenomena* (Clarendon Press, Oxford) 1993.
[19] RIORDAN O. and WORMALD N., *Combin. Comput.*, **19** (2010) 835.
[20] DEHMER M. and EMMERT-STREIB F., *Quantitative Graph Theory* (Chapman and Hall/CRC, Boca Raton, Fla., USA) 2014.
[21] WIENER H., *J. Am. Chem. Soc.*, **69** (1947) 17.
[22] MIHALIĆ Z. and TRINAJSTI N., *J. Chem. Educ.*, **69** (1992) 701.
[23] IVANCIUC O., BALABAN T. S. and BALABAN A. T., *J. Math. Chem.*, **12** (1993) 309.
[24] PLAVŠIĆ D., NIKOLIĆ S., TRINAJSTI N. and MIHALIĆ Z., *J. Math. Chem.*, **12** (1993) 235.
[25] CALDARELLI G., CAPOCCI A., DE LOS RIOS P. and MUÑOZ M. A., *Phys. Rev. Lett.*, **89** (2002) 258702.
[26] BOGUÑÁ M. and PASTOR-SATORRAS R., *Phys. Rev. E*, **68** (2003) 036112.