Leaf-to-leaf distances and their moments in finite and infinite $m$-ary tree graphs

A.M. Goldsborough, S.A. Rautu and R.A. Römer
Department of Physics and Centre for Scientific Computing, University of Warwick, Coventry, CV4 7AL, UK
E-mail: A.Goldsborough@warwick.ac.uk

Abstract. We study the leaf-to-leaf distances on full and complete $m$-ary graphs using a recursive approach. In our formulation, leaves are ordered along a line. We find explicit analytical formulae for the sum of all paths for arbitrary leaf-to-leaf distance $r$ as well as the average path lengths and the moments thereof. We show that the resulting explicit expressions can be recast in terms of Hurwitz-Lerch transcendants. Results for periodic trees are also given. For incomplete random binary trees, we provide first results by numerical techniques; we find a rapid drop of leaf-to-leaf distances for large $r$.

Revision: 1.29, compiled Tuesday 17th June, 2014, 01:26

PACS numbers: 02.10.Ox, 02.30.Gp

Submitted to: J. Phys. A: Math. Gen.
1. Introduction

The study of graphs and trees, i.e. objects (or *vertices*) with pairwise relations (or *edges*) between them, has a long and distinguished history throughout nearly all the sciences. In computer science, graphs, trees and their study are closely connected, e.g. with sorting and search algorithms [1]; in chemistry the Wiener number is a topological index intimately correlated with, e.g., chemical and physical properties of alkane molecules [2]. In physics, graphs are equally ubiquitous, not least because of their immediate usefulness for systematic perturbation calculations in quantum field theories [3]. In mathematics, graph theory is in itself an accepted branch of mainstream research and graphs are a central part of the field of discrete mathematics [4]. An important concept that appears in all these fields is the *distance* or *path length* in a graph, i.e. the distance between certain vertices, given in terms of the number of edges connecting them [5, 6, 7]. For trees, i.e. undirected graphs in which any two vertices are connected by one edge only, various results exist [8, 9, 10], for example, that compute the path lengths from the top of the tree to its final leaves. In a binary tree such as shown in Fig. 1 this path length might correspond, e.g. to the number of yes/no decisions one performs when searching for information.

Tree-like structures have recently also become more prominent in quantum physics with the advent of so-called tensor network methods [11]. These provide elegant and powerful tools for the simulation of quantum many-body systems. In a recent publication [12] we show that certain correlation functions and measures of quantum entanglement can be constructed by a holographic distance and connectivity dependence along a tree network connecting certain leaves [13]. In these quantum systems, the leaves are ordered according to their physical distance, for example the separation of magnetic ions in a quantum wire. This ordering imposes a new restriction on the tree itself and the path lengths which become important are leaf-to-leaf distances across the tree. In the present work, we shall concentrate on full and complete trees. We derive the average path lengths for varying leaf-to-leaf distances with leaves ordered in a one-dimensional

![Figure 1](image-url)

**Figure 1.** A complete binary tree with various definitions discussed in main text labelled. Circles (●, ○) denote vertices while lines indicate edges between the vertices of different depth. The tree as shown has a depth of 5 and \( L = 32 \) external nodes (○). The indicated separation \( r \) is 5 while the associated path lengths equals 8 as indicated by the bold line.
Figure 2. Schematic decomposition of a level $n$ tree with root node (●) and leaves (◦) into two level $n−1$ trees (rectangles) each of which contains $2^{n−1}$ leaves.

The method is then generalised to $m$-ary trees and the moments of the path lengths. Explicit analytical results are derived for finite and infinite trees. We also consider the case of periodic trees. Last, we numerically study the case of incomplete random graphs, which is closest related to the tree tensor networks considered in Ref. [12].

2. Average leaf-to-leaf path length in complete binary trees

2.1. Recursive formulation

Let us start by considering the complete binary tree shown in Figure 1. It is a connected graph where each vertex is 3-valent and there are no loops. The root node is the vertex with just two degrees at the top of Figure 1. The rest of the vertices each have two daughter nodes and one parent. A leaf node has no daughters. The depth of the tree denotes the number of vertices from the root node with the root node at depth zero. With these definitions, a binary graph is complete or perfect if all of the leaf nodes are at the same depth and all the levels are completely filled. We now denote by the level, $n$, a complete set of vertices that have the same depth. These are enumerated with the root level as 0. We will refer to a level $n$ tree as a complete tree where the leaves are at level $n$. The path length, $\ell$, is the number of edges that are passed to go from one external node to another (cp. Figure 1). We would like to bring attention to the fact that in some fields the path length refers to the sum of the levels of each of the vertices in the tree [1], whilst what we are studying is known as the distance [6].

Let us now impose an order on the tree of Figure 1 such that the external nodes are enumerated from left to right to indicate position values, $x_i$, for leaf $i$. Then we can define a leaf-to-leaf distance $r = |x_i − x_j|$ for any pair of leaves $i$ and $j$. This is equivalent to the notion of distance on a one-dimensional physical lattice. Let the length $L$ be the length of the lattice, i.e. number of external nodes. Then for such a complete binary tree, we have $L = 2^n$.

Clearly, there are many pairs of leaves are separated by $r$ from each other (cp. Figure 1). Let $\{\ell_n(r)\}$ denote the set of all corresponding path lengths. We now want to calculate the average path length $\mathcal{L}_n(r)$ from the set $\{\ell_n(r)\}$. We first note that for a level $n$ tree the number of possible paths with separation $r$ is given as $2^n − r$. In Figure 2 we see that any complete level $n$ tree can be decomposed into two level $n−1$ trees as shown e.g. in Fig. 1 for a binary tree [2].
sub-trees each of which contains \(2^{n-1}\) leaves. Let \(S_n(r)\) denote the sum of all possible path lengths encoded in the set \(\{\ell_n(r)\}\). The structure of the decomposition in Figure 2 suggests that we need to distinguish two classes of path lengths \(r\). First, for \(r < 2^{n-1}\), paths are either completely contained within each of the two level \(n-1\) trees or they bridge from the left level \(n-1\) tree to the right level \(n-1\) tree. Those which are completely contained sum to \(2S_{n-1}(r)\). For those path of length \(r\) that bridge across the two level \((n-1)\) trees, there are \(r\) of such paths and each path has lengths \(\ell_{n-1} = 2n\). Next, for \(r \geq 2^{n-1}\), paths no longer fit into a level \(n-1\) tree and always bridge from left to right. Again, each such path is \(2n\) long and there are \(L - r = 2^n - r\) such paths.

Putting it all together, we find that

\[
S_n(r) = \begin{cases} 
2S_{n-1}(r) + 2nr & r < 2^{n-1}, \\
2n(2^n - r) & r \geq 2^{n-1}.
\end{cases} 
\tag{1}
\]

for \(n > 1\) and with \(S_1(r) = 1\). Dividing by the total number of possible paths of length \(r\) then gives the desired average path length

\[
L_n(r) \equiv \frac{S_n(r)}{2^n - r}. 
\tag{2}
\]

2.2. An explicit expression

As long as \(r < 2^{n-1}\), equation (1) can be recursively expanded, i.e.

\[
S_n(r) = 2S_{n-1}(r) + 2nr 
= 2[2S_{n-2}(r) + 2(n-1)r] + 2nr 
= \ldots
\]

After \(\nu\) such expansions, we arrive at

\[
S_n(r) = 2^\nu S_{n-\nu}(r) + \sum_{k=0}^{\nu-1} 2^{k+1}(n-k)r. \tag{4}
\]

The expansion can continue while \(r < 2^{n-\nu-1}\). It terminates when \(n - \nu\) becomes so small such that the leaf-to-leaf distance \(r\) is no longer contained within the level-\((n-\nu)\) tree. Hence the smallest permissible value of \(n - \nu\) is given by

\[
n_c(r) = \lceil \log_2 r \rceil + 1, \tag{5}
\]

where \(\lceil \cdot \rceil\) denotes the floor function. For clarity, we will suppress the \(r\) dependence, i.e. we write \(n_c \equiv n_c(r)\) in the following. Continuing with the expansion of \(S_n(r)\) up to the \(n_c\) term, we find

\[
S_n(r) = 2^{n-n_c}S_{n_c}(r) + \sum_{k=0}^{n-n_c-1} 2^{k+1}(n-k)r
\]

\[
= 2^{n-n_c}S_{n_c}(r) + [2^{n-n_c+1}(n_c + 2) - 2(n+2)]r. \tag{6b}
\]

Details for the summations occurring in Equation (6b) are given in Appendix A. From Equation (1), we have \(S_{n_c}(r) = 2n_c(2^{n_c} - r)\). Thus Equation (6b) becomes

\[
S_n(r) = 2^{n+1}(n_c + 2^{1-n_c}r) - 2(n+2)r. \tag{7}
\]
Leaf-to-leaf distances in m-ary tree graphs

Figure 3. (a) The average path length $L_n(r)$ versus spatial distance $r$ for a complete binary tree of $n = 20$ (dashed), i.e. length $L = 2^{20} = 1,048,576$, and also for $n \to \infty$ (solid). The first 10 values are indicated by circles. (b) Average path length $L_n^{(m)}(r)$ for m-ary trees of various $m$. The curves for $m = 2, 5, 50$ are shown as solid lines, while those for $m = 3, 10$ and 100 have been indicated as dashed lines for clarity.

Hence the average path lengths are given by

$$L_n(r) = \frac{2}{2^n - r} \left[ 2^n (n_c + 2^{1-n_c}r) - (n+2)r \right]. \quad (8)$$

In the limit of $n \to \infty$ for fixed $r$, we have

$$\lim_{n \to \infty} L_n(r) \equiv L_\infty(r) = 2 \left( n_c + 2^{1-n_c}r \right). \quad (9)$$

We emphasise that $L_\infty(r) < \infty \forall r < \infty$. In Figure 3 we show finite and infinite path lengths $L_n(r)$. We see that whenever $r = 2^i$, $i \in \mathbb{N}$, we have a cusp in the $L_n(r)$ curves. Between these points, the $\lfloor \cdot \rfloor$ function enhances deviations from the leading $\log_2 r$ behavior. This behaviour is from the self-similar structure of the tree. Consider a sub-tree with $\nu$ levels, the largest separation that can occur in that sub-tree is $r = 2^{\nu}$, which has average length $2\nu$. When $r$ becomes larger than the sub-tree size the path length can no longer be $2\nu - 1$ but always larger, so there is a cusp where this path length is removed from the possibilities. The constant average length when $r \geq \frac{L_2}{2}$ is because there is only one possible path length that connects the two primary sub-trees, which is clear from (1).

3. Generalization to complete m-ary trees

3.1. Average leaf-to-leaf path length in complete ternary trees

Ternary trees are those where each node has three daughters. Let us denote by $S_n^{(3)}(r)$ and $L_n^{(3)}(r)$ the sum and average, respectively, of all possible path lengths $\{\ell_n^{(3)}(r)\}$ for given $r$ in analogy to the binary case discussed before. Furthermore, $L = 3^n$. Following the arguments which led to Equation (1), we have

$$S_n^{(3)}(r) = \begin{cases} 
3S_{n-1}^{(3)}(r) + 4nr & r < 3^{n-1}, \\
2(3^n - r) & r \geq 3^{n-1}.
\end{cases} \quad (10)$$
This recursive expression can again be understood readily when looking at the structure of a ternary tree. Clearly, $S^{(3)}_n(r)$ will now consist of the sum of path lengths for three level $n$ trees, plus the sum of all paths that connect the nodes across the three trees of level $n$. The lengths of these paths is solely determined by $n$ irrespective of the number of daughters and hence remains $2n$. As before, we need to distinguish between the case when $r$ fits within a level $n - 1$ tree, i.e. $r < 3^{n-1}$, and when it connects different level $n - 1$ trees, $r \geq 3^{n-1}$. For $r < 3^{n-1}$, there are now $2r$ such paths, i.e., $r$ between the left and centre level $n - 1$ trees and $r$ the centre and right level $n - 1$ trees. For $r \geq 3^{n-1}$ there are $L - r = 3^n - r$ paths. We again expand the recursion (10) and find, with $n_c^{(3)} = \lfloor \log_3 r \rfloor + 1$ in analogy to (5), that

$$S^{(3)}_n(r) = 3^n \left[ 2n_c^{(3)} + 3^{1-n_c^{(3)}} r \right] - (2n + 3)r$$

and

$$L^{(3)}_n(r) = \frac{S^{(3)}_n(r)}{3^n - r},$$

$$L^{(3)}_\infty(r) = 2n_c^{(3)} + 3^{1-n_c^{(3)}} r.$$  (13)

### 3.2. Average leaf-to-leaf path length in complete $m$-ary trees

The methodology and discussion of the binary and ternary trees can be generalised to trees of $m > 1$ daughters, known as $m$-ary trees. The maximal path length for any tree is independent of $m$ and determined entirely by the geometry of the tree. Each external node is at depth $n$, a maximal path has the root node as the lowest common ancestor, therefore the maximal path is $2n$.

A recursive function can be obtained using similar logic to before. For a given $n$, there are $m$ subgraphs with the structure of a tree with $n - 1$ levels. When $r$ is less than the size of each subgraph ($r < m^{n-1}$), the sum of the paths is therefore the sum of $m$ copies of the subgraph along with the paths that connect neighbouring pairs. When $r$ larger than the size of the subgraph ($r \geq m^{n-1}$), the paths are all maximal. When all this is taken into account the recursive function is

$$S^{(m)}_n(r) = \begin{cases} 
  mS^{(m)}_{n-1}(r) + 2(m - 1)nr & r < m^{n-1} \\
  2n(m^n - r) & r \geq m^{n-1}. 
\end{cases}$$

This can be solved in the same way as the binary case to obtain an expression for the sum of the paths for a given $m$, $n$ and $r$

$$S^{(m)}_n(r) = 2m^n \left[ n_c^{(m)} + \frac{m^{1-n_c^{(m)}}}{(m - 1)} \right] - 2r \left( n + \frac{m}{m - 1} \right),$$

The average path length is then

$$L^{(m)}_n(r) = \frac{S^{(m)}_n(r)}{m^n - r}.$$  (16)
and
\[ \mathcal{L}_\infty^{(m)}(r) = 2 \left( n_c^{(m)} + \frac{m^{1-n_c^{(m)}} r}{(m - 1)} \right). \]

We note that in analogy with Equation (5), we have used
\[ n_c^{(m)} = \lceil \log_m r \rceil + 1 \]
in deriving these expressions. Figure 3 shows the resulting path lengths in the \( n \to \infty \) limit for various values of \( m \).

4. Moments of the leaf-to-leaf path length distribution in complete \( m \)-ary trees

4.1. Variance of path lengths in complete \( m \)-ary trees

In addition to the average path length \( \mathcal{L}_n^{(m)}(r) \), it is also of interest to ascertain its variance \( \text{var}[\mathcal{L}_n^{(m)}](r) = \langle [\mathcal{L}_n^{(m)}(r)]^2 \rangle - [\mathcal{L}_n^{(m)}(r)]^2 \). Here \( \langle \cdot \rangle \) denotes the average over all paths for given \( r \) in an \( m \)-ary tree as before. In order to obtain the variance, we obviously need to obtain an expression for the sum of the squares of path lengths. This can again be done recursively, i.e. with \( \mathcal{Q}_n^{(m)}(r) \) denoting this sum of squared path length for an \( m \)-ary graph of leaf-to-leaf distance \( r \), we have similarly to Equation (14)
\[ \mathcal{Q}_n^{(m)}(r) = \begin{cases} m \mathcal{Q}_{n-1}^{(m)}(r) + (m - 1)4n^2 r & r < m^{n-1}, \\ 4n^2 (m^n - r) & r \geq m^{n-1}. \end{cases} \]

Here, the difference to Equation (14) is that we have squared the length terms \( 2n \). As before, expanding down to \( n_c \) (here and in the following, we suppress the \( (m) \) superscript of \( n_c^{(m)} \) for clarity) gives a term containing \( \mathcal{Q}_{n_c}^{(m)}(r) \),
\[ \mathcal{Q}_n^{(m)}(r) = m^{n-n_c} \mathcal{Q}_{n_c}^{(m)}(r) + \sum_{k=0}^{n-n_c-1} 4(m-1)(n-k)^2 m^k r 
= m^{n-n_c} \mathcal{Q}_{n_c}^{(m)}(r) + 4r(m-1) \sum_{k=0}^{n-n_c-1} [n^2 m^k - 2nkm^k + k^2 m^k] 
= \frac{4}{(m-1)^2} \left\{ r m^{n-n_c+1} [m + 2n_c(m-1) + 1] + m^n (m-1)^2 n_c^2 
- r \left[ n^2 + m^2(n+1)^2 + m(1-2n(n+1)) \right]\right\}. \]

As before, details for the summations occurring in Equation (20b) are given in Appendix A. We can therefore write for the variance
\[ \text{var}[\mathcal{L}_n^{(m)}](r) = \frac{\mathcal{Q}_n^{(m)}(r)}{m^n - r} - \left[ \mathcal{L}_n^{(m)}(r) \right]^2 = \frac{\mathcal{Q}_n^{(m)}(r)}{m^n - r} - \left[ \frac{\mathcal{S}_n^{(m)}(r)}{m^n - r} \right]^2. \]

Using Equation (20c), (16) and (15), we then have explicitly
\[ \text{var}[\mathcal{L}_n^{(m)}](r) = \frac{4r}{m^{2n_c-2}(m^n - r)^2(m-1)^2} \left[ m^{2n_c-1}(m+1) - r \right] + m^{2n_c-1} r - \]
By expanding, this gives

\[ r = \frac{4r [m^{n-1}(m + 1) - 1]}{m^{2n-2}(m - 1)^2} \]  

and also

\[ \text{var}[\mathcal{L}_n^{(m)}](r) = \frac{4r [m^{n-1}(m + 1) - 1]}{m^{2n-2}(m - 1)^2}. \]

When \( r = m^i, i \in \mathbb{N}^0 \), then \( \text{var}[\mathcal{L}_\infty^{(m)}] \) has a local minima and we find that \( \text{var}[\mathcal{L}_\infty^{(m)}](m^i) = \frac{4m^{2n-2}}{(m-1)^2} \). Similarly, it can be shown that the local maxima are at \( r = \frac{1}{2}m^i(m + 1) \), then \( \text{var}[\mathcal{L}_\infty^{(m)}] = \frac{4m^{2n-2}}{(m-1)^2} + 1 \). These values are indicated in Figure 4 for selected \( m \).

4.2. General moments of path lengths in complete \( m \)-ary trees

The derivation in section 4.1 suggests that any \( q \)-th raw moment of path lengths can be calculated similarly as in Equation (20). Indeed, let us define \( \mathcal{M}_{q,n}^{(m)}(r) \) as the \( q \)-th moment of an \( m \)-ary tree of level \( n \) with leaf-to-leaf distance \( r \). Then \( \mathcal{M}_{1,n}^{(m)}(r) = \mathcal{L}_n^{(m)}(r) \), \( \mathcal{M}_{2,n}^{(m)}(r) = \mathcal{Q}_n^{(m)}(r) \) and \( \text{var}[\mathcal{L}_n^{(m)}](r) = \frac{\mathcal{M}_{2,n}^{(m)}(r)}{m^{n-r}} - \left[ \frac{\mathcal{M}_{1,n}^{(m)}(r)}{m^{n-r}} \right]^2 \). Following Equation (19), we find

\[ \mathcal{M}_{q,n}^{(m)}(r) = \begin{cases}  
  m\mathcal{M}_{q,n-1}^{(m)}(r) + 2^q n^q (m - 1)r & r < m^{n-1}, \\
  2^q n^q (m^n - r) & r \geq m^{n-1}.
\end{cases} \]  

By expanding, this gives

\[ \mathcal{M}_{q,n}^{(m)}(r) = m^{n-n_c} \mathcal{M}_{q,n_c}^{(m)}(r) + \sum_{k=0}^{n-n_c-1} 2^q m^k (m-1)(n-k)^q r. \]
As before, \( n_c \) corresponds to the first \( n \) value where, for given \( r \), we have to use the second part of the expansion as in Equation (24). Hence we can substitute the second part of (25) for \( \mathcal{M}_{q,n}(r) \) giving

\[
\mathcal{M}_{q,n}(r) = m^{n-n_c}2^q n_c^q (m^{n_c} - r) + \sum_{k=0}^{n-n_c-1} 2^q m^k (m - 1)(n - k)^q r.
\]  

(26)

In order to derive and explicit expression for this similar to section 2.2, we need again to study the final sum of Equation (26). We write

\[
\sum_{k=0}^{n-n_c-1} 2^q m^k (m - 1)(n - k)^q r = r(m - 1)(-2)^q \left[ \sum_{k=0}^{\infty} m^k (k - n)^q - \sum_{k=n-n_c}^{\infty} m^k (k - n)^q \right]
\]  

(27a)

\[
= r(m - 1)(-2)^q \left[ \sum_{k=0}^{\infty} m^k (k - n)^q - m^{n-n_c} \sum_{k=0}^{\infty} m^k (k - n_c)^q \right]
\]  

(27b)

\[
= r(m - 1)(-2)^q \left[ \Phi(m, -q, -n) - m^{n-n_c} \Phi(m, -q, -n_c) \right]
\]  

(27c)

where in the last step we have introduced the Hurwitz-Lerch Zeta function \( \Phi \) (also referred to as the Lerch transcendent or the Hurwitz-Lerch Transcendent). It is defined as the sum

\[
\Phi(z, s, u) = \sum_{k=0}^{\infty} \frac{z^k}{(k + u)^s}, \quad z \in \mathbb{C}.
\]  

(28)

The properties of \( \Phi(z, s, u) \) are

\[
\Phi(z, s, u + 1) = \frac{1}{z} \left( \Phi(z, s, u) - \frac{1}{u^s} \right),
\]  

(29a)

\[
\Phi(z, s - 1, u) = \left( u + z \frac{\partial}{\partial z} \right) \Phi(z, s, u),
\]  

(29b)

\[
\Phi(z, s + 1, u) = - \frac{1}{s} \frac{\partial \Phi}{\partial u}(z, s, u).
\]  

(29c)

Hence we can write

\[
\mathcal{M}_{q,n}(r) = m^{n-n_c}2^q n_c^q (m^{n_c} - r) + r(m - 1)(-2)^q \left[ \Phi(m, -q, -n) - m^{n-n_c} \Phi(m, -q, -n_c) \right].
\]  

(30)

Averages of \( \mathcal{M}_{q,n}(r) \) can be defined as previously via

\[
\mathcal{A}_{q,n}(r) = \frac{\mathcal{M}_{q,n}(r)}{m^n - r}
\]  

(31)

such that \( \mathcal{L}_n(r) = \mathcal{A}_{1,n}(r) \) and \( \text{var}[\mathcal{L}_n(r)](r) = \mathcal{A}_{2,n}(r) - \left[ \mathcal{A}_{1,n}(r) \right]^2. \)
The properties (29a) – (29c) can be used to show that, for a given $m$ and $q$, $\Phi (m, -q, -n)$ can be expressed as a polynomial of order $(-n)^q$. Therefore in the $n \to \infty$ limit, we find

$$\lim_{n \to \infty} \mathcal{A}^{(m)}_{q,n}(r) \equiv \mathcal{A}^{(m)}_{q,\infty}(r) = m^{-n_c} [2^n n_c^q (m^{n_c} - r) - r(m - 1)(-2)^q \Phi (m, -q, -n_c)].$$

(32)

5. Complete $m$-ary trees with periodicity

Up to now we have always dealt with trees in which the maximum distance $r$ was set by the number of leaves, i.e. $r \leq m^n$. This is known as a hard wall or open boundary in terms of physical systems defined along $r$. A periodic boundary can be realised by having the leaves of the tree form a circle as depicted in Figure 5 for a binary tree. For such a binary tree, only distances $r \leq L/2$ are relevant since all cases with $r > L/2$ can be reduced to smaller $r = \text{mod}(r, L/2)$ values by going around the periodic tree in the opposite direction. Therefore we can write

$$\mathcal{M}^{(m,\circ)}_{1,n}(r) = \mathcal{M}^{(m)}_{1,n}(r) + \mathcal{M}^{(m)}_{1,n}(m^n - r),$$

(33)

where $r < L/2$ and the subscript $\circ$ denotes the periodic case. Note that the case where $r = L/2$ the clockwise and anti-clockwise paths are the same so only need to be counted once. In the simple binary tree case we can expand this via (7) as in section 2.2 and find

$$\mathcal{M}^{(2,\circ)}_{1,n}(r) \equiv \mathcal{S}^{(2,\circ)}_{n}(r) = 2^{n+1} \left[ n_c + \tilde{n}_c - n - 2 + 2^{1-n_c} r + 2^{1-\tilde{n}_c} (2^n - r) \right],$$

(34)
with \( n_c \) as in Equation (5) and \( \tilde{n}_c = \lfloor \log_2(2^n - r) \rfloor + 1 \). For every \( r \), we have \( 2^n \) possible starting leaf positions on a periodic binary tree and hence the average path length can be written as

\[
\mathcal{A}^{(2,0)}_{1,n}(r) \equiv \mathcal{L}^{(2,0)}_{n}(r) = \frac{S^{(2,0)}_{n}(r)}{2^n} = 2 \left[ n_c + \tilde{n}_c - n - 2 + 2^{1-n_c}r + 2^{1-\tilde{n}_c}(2^n - r) \right]. \tag{35}
\]

This expression is the periodic analogue to Equation (8). Generalizing to \( m \)-ary trees, with \( \tilde{n}_c = \lfloor \log_m(m^n - r) \rfloor + 1 \), we find

\[
\mathcal{M}^{(m,0)}_{1,n}(r) = \mathcal{M}^{(m)}_{1,n}(r) + \mathcal{M}^{(m)}_{1,n}(m^n - r)
= 2m^n \left[ n_c + \tilde{n}_c - n - 2 + \frac{1}{m-1} \left( m^{1-n_c}r + m^{1-\tilde{n}_c}(m^n - r) - m \right) \right]. \tag{36}
\]

The average path length for \( m \)-ary periodic trees is then given as

\[
\mathcal{A}^{(m,0)}_{1,n}(r) = \frac{\mathcal{M}^{(m,0)}_{1,n}(r)}{m^n} = 2 \left[ n_c + \tilde{n}_c - n - 2 + \frac{1}{m-1} \left( m^{1-n_c}r + m^{1-\tilde{n}_c}(m^n - r) - m \right) \right]. \tag{38}
\]

To again study the case of \( n \to \infty \), it is necessary to observe how \( \tilde{n}_c \) behaves for large \( n \) and fixed \( m, r \). When \( n \gg r \), we have \( r < m^{n-1} \) and hence \( \lim_{n \to \infty} [\log_m(m^n - r)] = n - 1 \). This enables us to simply take the limits of Equation (38) to give

\[
\lim_{n \to \infty} \mathcal{A}^{(m,0)}_{1,n}(r) \equiv \mathcal{A}^{(m,0)}_{1,\infty}(r) = 2 \left[ n_c + \frac{m^{1-n_c}r}{(m-1)} \right], \tag{39}
\]

which is the same as the open boundary case (17). This is to be expected as a small region of a large circle can be approximated by a straight line.

Last, the \( q \)-moments can be expressed similarly to Equation (30) via the Lerch transcendent as

\[
\mathcal{M}^{(m,0)}_{q,n}(r) = \mathcal{M}^{(m,q)}_{q,n}(r) + \mathcal{M}^{(m)}_{q,n}(m^n - r), \tag{40}
\]

\[
= m^{n-n_c}2^q\tilde{n}_c^q(m^{n_c} - r) + m^{n-n_c}2^q \tilde{n}_c^q(m^{\tilde{n}_c} - m^n + r) + (m - 1)(-2)^q \left[ m^n\Phi(m, -q, -n) - rm^{n-n_c}\Phi(m, -q, -n_c) \right.
- \left. (m^n - r)m^n\tilde{n}_c\Phi(m, -q, -\tilde{n}_c) \right]. \tag{41}
\]

The average \( q \)-moments in full are therefore

\[
\mathcal{A}^{(m,0)}_{q,n}(r) = \frac{\mathcal{M}^{(m,0)}_{q,n}(r)}{m^n} \tag{42}
\]

for a complete, periodic, \( m \)-ary tree. To take the limit \( n \to \infty \) notice that \( \tilde{n}_c = n \) when \( r < m^{n-1} \) for large \( n \). Just like with Equation (39), this results in \( \mathcal{A}^{(m,0)}_{q,\infty}(r) = \mathcal{A}^{(m)}_{q,\infty}(r) \).

6. Path lengths for random binary trees

In Figure 6 we show a binary graph where the leaves do not all appear at the same level \( n \), but rather each node can become a leaf node according to an independent and identically distributed random process. Such graphs are no longer complete, but nevertheless have many applications in the sciences [11, 12]. Let us again compute the average path length.
Leaf-to-leaf distances in m-ary tree graphs

Figure 6. (a) A random binary tree. (b) A complete set of random binary trees for $n = 1, 2, 3$ (L = 2, 3, 4). Circles and lines are as in Figure 1.

Figure 7. (a) Average path length through a random binary graph connecting two leaves of separation $r$ averaged over all possible graphs for $L = 9, 10$ and 11 (solid symbols, lines are guide to the eye only). The open symbols (dashed line guide to the eye) refer to an average over 10000 randomly chosen graphs from the 10! possibilities for $L = 11$. The grey crosses ($\times$) and line correspond to $L^{(2)}(r)$ from Equation (9). (b) Average path length constructed from 500 randomly chosen binary trees with $L = 1000$ (dashed line). The open symbols (◦) denotes the first 10 data points. The closed symbols (red •) and the solid line correspond to the $L = 10$ data from (a). The grey line correspond to $L^{(2)}(r)$ as in (a). Error bars have been omitted in (a) and (b) as they are within symbol size.

$L^{(2, \mathcal{R})}(r)$ for a given $r$, when all possible pairs of leaves of distance $r$ and all possible trees of $L - 1$ internal nodes are considered. Here $\mathcal{R}$ denotes the random character of trees under consideration. For each $n = L - 1$, there are $n!$ different such random graphs as shown in Figure 6. We construct these graphs numerically and measure $L^{(2, \mathcal{R})}(r)$ as shown in Figure 7. For small $n$, we have computed all $(L - 1)!$ graphs (cp. Figure 7(a)) while for large $n$, we have averaged over a finite number $N \ll (L - 1)!$ of randomly chosen binary trees among the $(L - 1)!$ possible trees (cp. Figure 7(b)). We see in Figure 7(a)§ we emphasise that this definition of a random graph is different from the definition of so-called Catalan tree graphs [1], as the number of unique graphs is given by the Catalan number $C_n$ and does not double count the degenerate graphs as shown in Figure 6.

§
that, similar to the complete binary trees considered in the section, the path lengths increase with $r$ until they reach a maximal value. Then they start to decrease rapidly unlike the complete graph in Figure 3. We also see that for such small trees, we are still far from the infinite complete tree result $\mathcal{L}_\infty^{(2)}(r)$ of Equation (9). Finally, we also see that when we choose 10,000 random binary trees from the $10! = 3,628,800$ possible such trees at $L = 11$ that the average path lengths for each $r$ is still distinguishably different from an exact summation of all path lengths. This suggests that rare tree structures are quite important. In Figure 7(b) we nevertheless show estimates of $\mathcal{L}_n^{(2,\mathcal{R})}(r)$ for various $n$. As before, the shape of the curves for large $n$ is similar to those for small $n$. Clearly, however, the cusps in $\mathcal{L}_n^{(2)}(r)$ are no longer present in $\mathcal{L}_n^{(2,\mathcal{R})}(r)$. Also, the values of $\mathcal{L}_n^{(2,\mathcal{R})}(r)$ are larger than those for $\mathcal{L}_n^{(2)}(r)$ for small $r$.

7. Conclusions

We have calculated an analytic form for the average length of the path that separates two leaves with a given separation — ordered according to the physical distance on a line — in a complete binary tree graph. This result is then generalised to a complete tree where each vertex has any number of children. In addition to the mean path length, it is found that the raw moments of the distribution of path lengths have an analytic form that can be expressed in a concise way in terms of the Hurwitz-Lerch Zeta function. These findings are calculated for open trees, where the leaves form an open line, periodic trees, where the leaves form a circle, and infinite trees, which is the limit where the number of levels, $n$, goes to infinity. Each of these results has a neat form and characteristic features due to the self-similarity of the trees.

Acknowledgments

We thank M. Bates, A. Czumaj and G. Rowlands for discussions. We would like to thank the EPSRC for financial support (EP/J003476/1) and provision of computing resources through the MidPlus Regional HPC Centre (EP/K000128/1).

Appendix A. Some useful series expressions

Appendix A.1. Series used in section 2.2

When the last sum in Equation (6a) is expanded, it is simply the sum of two geometric series. The first part can be simplified using

$$\sum_{k=1}^{l} x^k = \frac{x(1-x^l)}{1-x}, \quad (A.1)$$

the second part with

$$\sum_{k=1}^{l} kx^{k+1} = \frac{x(1-x^{l+1})}{(1-x)^2} = \frac{x + lx^{l+2}}{1-x}, \quad (A.2)$$
Leaf-to-leaf distances in $m$-ary tree graphs

and we also use

$$
\sum_{k=0}^{l} x^k = \frac{1 - x^{l+1}}{1 - x}. \quad (A.3)
$$

Appendix A.2. Series used in section 4.1

The explicit expressions for the series terms occurring in Equation (20) are given here. The first part is a simple geometric series given by equation (A.3). The second part is an arithmetico-geometric series similar to (A.2).

$$
\sum_{k=0}^{l} kx^k = \frac{x(1 - x^l)}{(1 - x)^2} - \frac{lx^{l+1}}{1 - x}. \quad (A.4)
$$

The final part is another also an arithmetico-geometric series and has the following form [18]:

$$
\sum_{k=0}^{l-1} k^2 x^k = \frac{1}{(1 - x)^3} \left[ (-l + 2l - 1)x^{l+2} + (2l^2 - 2l - 1)x^{l+1} - l^2 x^l + x^2 + x \right]. \quad (A.5)
$$

References

[1] Robert Sedgewick and Philippe Flajolet. *An Introduction to the Analysis of Algorithms*. Addison-Wesley, Westford, MA, second edition, 2013.

[2] Harry Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1):17–20, 1947.

[3] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to Quantum Field Theory*. Westview Press, Boulder, CO, 1995.

[4] Kenneth H. Rosen. *Discrete Mathematics and its Applications*. McGraw Hill, New York, NY, seventh edition, 2012.

[5] Kevin Leckey, Ralph Neininger, and Wojciech Szpankowski. Towards more realistic probabilistic models for data structures: The external path length in tries under the Markov model. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2013.

[6] L. A. Szekely, Hua Wang, and Taoyang Wu. The sum of the distances between the leaves of a tree and then ‘semi-regular’ property. *Discrete Mathematics*, 311:1197–1203, 2011.

[7] Hua Wang. Sum of distances between vertices/leaves in $k$-ary trees. *ICA bulletin*.

[8] Philippe Jacquet and Mireille Regnier. Normal limiting distribution for the size and the external path length of tries. Technical Report RR-0827, INRIA, April 1988.

[9] Peter Kirschenhofer, Helmut Prodinger, and Wojciech Szpankowski. On the balance property of Patricia tries: External path length viewpoint. *Theoretical Computer Science*, 68(1):1 – 17, 1989.

[10] P. Kirschenhofer, H. Prodinger, and W. Szpankowski. Digital search trees again revisited: The internal path length perspective. *SIAM Journal on Computing*, 23(3):598–616, 1994.

[11] Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Ann. Phys.*, 326(1):96–192, 2011.

[12] Andrew M. Goldsborough and Rudolf A. Römer. Self-assembling tensor networks and holography in disordered spin chains. *Phys. Rev. B*, 2014. accepted for publication, arXiv:1401.4874.

[13] G. Evenbly and G. Vidal. Tensor Network States and Geometry. *J. Stat. Phys.*, 145(4):891–918, 2011.

[14] S. Kanemitsu, M. Katsurada, and M. Yoshimoto. On the Hurwitz-Lerch zeta-function. *aequationes mathematicae*, 59(1-2):1–19, 2000.
[15] H.M. Srivastava. Riemann, Hurwitz and Hurwitz-Lerch Zeta functions and associated series and integrals. In Panos M. Pardalos and Themistocles M. Rassias, editors, Essays in Mathematics and its Applications, pages 431–461. Springer Berlin Heidelberg, 2012.

[16] Jesús Guillera and Jonathan Sondow. Double integrals and infinite products for some classical constants via analytic continuations of Lerch’s transcendent. The Ramanujan Journal, 16(3):247–270, 2008.

[17] Wolfram Research Inc. Mathematica Version 9.0. Champaign, IL, 2012.

[18] I. S. Gradshteyn and I. M. Ryzhik. Table of Integrals, Series and Products. Academic Press, London, 1994.