Journal of Integer Sequences, Vol. 17 (2014),
Article 14.6.3

On the Average Path Length of Complete m-ary Trees

M. A. Nyblom
School of Mathematics and Geospatial Science
RMIT University
Melbourne, Victoria 3001
Australia
michael.nyblom@rmit.edu.au

Abstract

Define the average path length in a connected graph as the sum of the length of the shortest path between all pairs of nodes, divided by the total number of pairs of nodes. Letting S_N denote the sum of the shortest path lengths between all pairs of nodes in a complete m-ary tree of depth N, we derive a first-order linear but non-homogeneous recurrence relation for S_N, from which a closed-form expression for S_N is obtained. Using this explicit expression for S_N, we show that the average path length within this graph/network is asymptotic to $D - \frac{4}{m-1}$, where D is the diameter of the m-ary tree, that is, the longest shortest path. This asymptotic estimate for the average path length confirms a conjectured asymptotic estimate in the case of complete binary tree.

1 Introduction

A network is a graph $G = (E, V)$ having no loops, which is also connected in that one can find a sequence of edges in E, or path, connecting any two pairs of vertices in V. The study of networks has in recent times blossomed into a new field which has found applications in such diverse areas as communication engineering and sociology. One focus of attention in this new discipline area, coined "Network Science" [3], is the problem of designing networks in an economic way, whereby a sequence of edges a path within the network can be constructed to connect any two pairs of nodes, while keeping the total number of edges used to a minimum. Such highly edge-efficient networks are useful in the area of designing communication networks, as they help reduce transmission delays. One way to compare the edge-efficiency of different networks $G = (E, V)$, is to introduce a new metric [3, p. 74], known as the edge efficiency of G, and defined by

$$E(G) = 1 - \frac{\bar{P}(G)}{|E|} \ ,$$

where $\bar{P}(G)$ is the average path length of G, which is calculated as the sum of the length of the shortest paths connecting all pairs of nodes in V, divided by the total number of pairs of nodes in V. (Note if there is more than one shortest path connecting a pair of nodes, then the length of these paths will still occur once as a summand used in the calculation of $\bar{P}(G)$.)

One can view $E(G)$ as a measure of how effectively edges are used within a network to minimize the average path length. As $\bar{P}(G) \leq |E|$ we see that $E(G)$ ranges from 0 when $\bar{P}(G) = |E|$, to approximately 1 when $\bar{P}(G) \cong 0$. The case $E(G) = 0$ corresponds to a worst case when the average path length is as large as possible, while if $E(G) \cong 1$ then the average path length is small relative to the number of edges, and so every edge can be seen as contributing to the edge efficiency of the network. A network in which $E(G)$ approaches 1 as $|E|$ approaches infinity has been described as being "scalable" [3, p. 74]. To experimentally demostrate this scalable property of such regular networks as the complete binary trees, T. Lewis [3, p. 76] developed a breadth-first algorithm for the computation of $\bar{P}(G)$, for various increasing depths N of these binary trees. As a consequence of his experimental results, it was conjectured that for a complete binary tree $\bar{P}(G) \sim D - 4$, as $N \to \infty$, where D is the diameter of the tree, that is, the longest shortest path corresponding to the value of N. It should be noted that the diameter is a frequently used characteristic of trees within the literature [4].

In this paper we shall prove this asymptotic estimate by establishing a more general result, namely that $\bar{P}(G) \sim D - \frac{4}{m-1}$ as $N \to \infty$, in the case when $G = (E, V)$ is a complete m-ary tree, that is, a rooted tree in which all nodes have exactly m children, with the exception of the leaf nodes that have no children, and which are all located at the highest depth (see Figure 1). To achieve this end, we first shall derive a closed-form expression for the sum, denoted S_N, of the shortest path lengths between all pairs of nodes in a complete m-ary tree of depth N. As a consequence of this expression in (1), it will then be an easy task to show that $\bar{P}(G) = \frac{S_N}{\binom{|V|}{2}} \sim D - \frac{4}{m-1}$ as $N \to \infty$. It should be noted that the quantity measured by S_N, is also referred to as a *Wiener index* in the area of computational chemistry, where graphs such as rooted trees represent the molecular graphs of certain chemical compounds. Although there is a well-known recursive procedure for the calculation of the Wiener index of such graphs [1], it has been observed [2] that this general procedure can be cumbersome to apply for particular families of rooted trees having a regular graph structure, such as the complete m-ary trees. It is for this reason we have in this paper directly exploited the graph structure of the m-ary tree, to produce a first-order linear but non-homogeneous recurrence relation for the calculation of S_N.

2 Main result

To help establish the main result we begin by making two simple observations concerning the complete m-ary tree. With reference to Figure 1, if the schematic diagram represents a complete m-ary tree of depth $N+1$ in which the root node is at depth 1, then note that the m subtrees whose root nodes are located at depth 2 must also be complete m-ary trees of depth N. Consequently one should be able to express S_{N+1} in terms of S_N, so establishing a recurrence relation for the sequence $\{S_N\}$ for $N \geq 2$. Secondly observe that as each node, with the exception of the root node, has a unique parent node located at the previous depth, there can only be one shortest path connecting any pair of nodes in a complete m-ary tree. It should be noted that uniqueness of the shortest path is a property enjoyed by all trees.

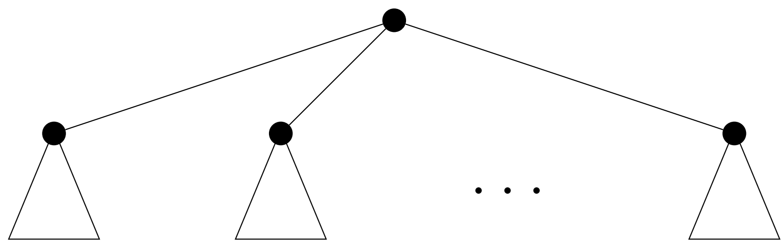

Figure 1: An m-ary tree of depth $N+1$

We shall also need the following elementary technical lemma.

Lemma 1. *For any $x \in \mathbb{R}\backslash\{1\}$ and integer $N \geq 1$*

$$\sum_{n=1}^{N} nx^{n-1} = \frac{1}{(x-1)^2}(Nx^{N+1} - Nx^N - x^N + 1) \ .$$

Proof. Differentiate both sides of the identity $1 + x + x^2 + \cdots x^N = (x^{N+1} - 1)/(x-1)$ with respect to x. □

By exploiting the above simple facts, one can now obtain a closed-form expression for S_N as follows.

Proposition 2. *If $G = (E, V)$ is a complete m-ary tree of depth $N \geq 2$, then the sum of the shortest path lengths between all pairs of nodes in V is given by*

$$S_N = m^{2N}\left(\frac{N}{(m-1)^2} - \frac{m+1}{(m-1)^3}\right) + m^N\left(\frac{N}{(m-1)^2} + \frac{m+1}{(m-1)^3}\right) \ . \tag{1}$$

Proof. The argument used to establish (1) will be broken into two parts. In Part 1 we use the geometry of the complete m-ary tree to derive a first-order linear but non-homogeneous recurrence relation for S_N. Although the process of solving such recurrence relations is standard a sketch of the details leading towards (1) is outlined in Part 2, for completeness.

Part 1: Again with reference to Figure 1, observe that the set S of shortest paths connecting pairs of nodes in a complete m-ary tree of depth $N+1$, can be partitioned as $S = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$, where $\mathcal{S}_1$ is the set of shortest paths connecting pairs of nodes located in any one but only one of the m subtrees of depth N, $\mathcal{S}_2$ is the set of shortest paths connecting any node contained in the m subtrees of depth N to the root of the tree of depth $N+1$, and finally $\mathcal{S}_3$ is the set of shortest paths connecting pairs of nodes located in any two separate subtrees of depth N. Clearly by definition, the sum of the lengths of the elements in $\mathcal{S}_1$ must be mS_N. If we denote the sum of the lengths of the elements in $\mathcal{S}_2$ and $\mathcal{S}_3$ by X_N and Y_N respectively, we can conclude that

$$S_{N+1} = mS_N + X_N + Y_N \ , \tag{2}$$

for $N \geq 2$. We now determine explicit expressions for X_N and Y_N, in terms of m and N. Considering any one of the m subtrees of depth N in Figure 1, recall from definition that the total number of nodes found at a depth of n, were $1 \leq n \leq N$, is m^{n-1}. Now in general from the geometry of a complete m-ary tree, observe that the length of the shortest path connecting any node at depth say i, to the root of the tree must be $i-1$. As there is only one edge connecting the root of the m subtrees, to the root of the tree of Figure 1, we conclude that the length of the shortest path connecting a node at a depth n in any subtree, to the root of the tree is $(n-1)+1 = n$. Consequently the sum of the shortest path lengths connecting all nodes in any one subtree of depth N, to the root of the tree is $\sum_{n=1}^{N} nm^{n-1}$. Thus by Lemma 1, one concludes after setting $x = m$ that

$$X_N = m\sum_{n=1}^{N} nm^{n-1} = \frac{m}{(m-1)^2}(Nm^{N+1} - Nm^N - m^N + 1) \ .$$

To determine Y_N, we first must show that the sum of the lengths of the shortest paths connecting any one node of a given subtree, at a depth $1 \leq n \leq N$, to all nodes located in a separate subtree of Figure 1, is given by

$$S(n) = (n+1)m^0 + (n+2)m^1 + (n+3)m^2 + \cdots + (n+N)m^{N-1} \ . \tag{3}$$

To this end, recall that the length of the shortest path from a node at depth $1 \leq n \leq N$ in any subtree, to the root of the tree is n. Consequently a shortest path beginning at a node at depth $1 \leq n \leq N$ in one subtree, to a node at depth $1 \leq i \leq N$ in another subtree, must first traverse a path of length n from the initial subtree to the root, then a path of length 1 to the root of the other subtree, and finally a path of length $i-1$ from this root to the node at depth $1 \leq i \leq N$ in the second subtree. Thus the length of such a shortest a path is $n+1+i-1 = n+i$, and as there are m^{i-1} nodes in the second subtree at depth i, we can conclude the sum of these individual shortest path lengths starting at a node at depth n in one subtree, and terminating at the nodes at depth i in another subtree is $(n+i)m^{i-1}$. Finally by adding the terms $(n+i)m^{i-1}$ over all depths $1 \leq i \leq N$ in the second subtree, one arrives at the expression in (3).

Now as there are m^{n-1} nodes at a depth n in any one subtree of Figure 1, observe that $\sum_{n=1}^{N} m^{n-1}S(n)$ must represent the total sum of the shortest path lengths connecting pairs of nodes located in any two separate subtrees of depth N. Hence by choosing the m subtrees two at a time, we deduce from definition that $Y_N = \binom{m}{2} \sum_{n=1}^{N} m^{n-1} S(n)$. To determine Y_N explicitly first write $S(n)$ as follows

$$S(n) = \sum_{i=1}^{N} (n+i) m^{i-1} = n \sum_{i=1}^{N} m^{i-1} + \sum_{i=1}^{N} i m^{i-1} = n \frac{m^N - 1}{m-1} + \sum_{i=1}^{N} i m^{i-1} \ . \tag{4}$$

Then after substituting the right hand side of (4) into the above expression for Y_N, one finds after an application of Lemma 1 that

$$\begin{aligned} Y_N &= \binom{m}{2} \sum_{n=1}^{N} m^{n-1} \left(n \left(\frac{m^N-1}{m-1} \right) + \sum_{i=1}^{N} i m^{i-1} \right) \\ &= \binom{m}{2} \left(\left(\frac{m^N-1}{m-1} \right) \sum_{n=1}^{N} n m^{n-1} + \sum_{n=1}^{N} m^{n-1} \sum_{i=1}^{N} i m^{i-1} \right) \\ &= \binom{m}{2} \left(\left(\frac{m^N-1}{m-1} \right) \sum_{n=1}^{N} n m^{n-1} + \left(\frac{m^N-1}{m-1} \right) \sum_{i=1}^{N} i m^{i-1} \right) \\ &= 2\binom{m}{2} \left(\frac{m^N-1}{m-1} \right) \sum_{n=1}^{N} n m^{n-1} \\ &= (m^{N+1} - m) \left(\frac{N m^{N+1} - N m^N - m^N + 1}{(m-1)^2} \right) \ . \end{aligned}$$

Adding these explicit expressions for X_N and Y_N and substituting the result into (2), produces the required recurrence relation for S_N as follows

$$S_{N+1} - m S_N = \left(\frac{m}{m-1} N - \frac{m}{(m-1)^2} \right) m^{2N} + \frac{m}{(m-1)^2} m^N \ , \tag{5}$$

for $N \geq 2$. The initial condition is clearly given by $S_2 = m + 2\binom{m}{2} = m^2$.

Part 2: As the recurrence relation in (5) is linear, recall that the solution S_N must be of the form $S_N = A m^N + B N m^N + (CN + D) m^{2N}$, where $A m^N$ corresponds to a general homogeneous solution of (5), and the constants B, C, D will be chosen so that the terms $B N m^N$ and $(CN + D) m^{2N}$ satisfy (5), with a corresponding right-hand side equal to $\frac{m}{(m-1)^2} m^N$ and $(\frac{m}{m-1} N - \frac{m}{(m-1)^2}) m^{2N}$ respectively. A standard calculation reveals that $B = C = \frac{1}{(m-1)^2}$ and $D = -\frac{m+1}{(m-1)^3}$ and so

$$S_N = A m^N + \frac{N}{(m-1)^2} m^N + \left(\frac{N}{(m-1)^2} - \frac{m+1}{(m-1)^3} \right) m^{2N} \ .$$

Applying the initial condition $S_2 = m^2$ one finally deduces that $A = \frac{m+1}{(m-1)^3}$, which results in the closed-form expression in (1). □

To state and prove the main result, recall that the diameter D of a network $G = (E, V)$, is the length of the longest shortest path connecting two nodes in V. In the case of a complete m-ary tree of depth $N \geq 2$, the diameter must be the length of the path connecting the left-most leaf node to the right-most leaf node of the tree, and so $D = 2(N - 1)$.

Theorem 3. *For a network $G = (E, V)$ which is a complete m-ary tree, the average path length $\bar{P}(G)$ satisfies the following asymptotic estimate*

$$\bar{P}(G) \sim D - \frac{4}{m-1} \quad \textit{as} \quad N \to \infty \ .$$

Proof. In a complete m-ary tree of depth $N \geq 2$ the total number of nodes is $|V| = 1 + m + m^2 + \cdots + m^{N-1} = (m^N - 1)/(m - 1)$. Recall the average path length of $G = (E, V)$ at a depth $N \geq 2$ is given by $\bar{P}(G) = \frac{S_N}{\binom{|V|}{2}}$ and as $\binom{|V|}{2} = O(m^{2N})$ we deduce from (1) that

$$\begin{aligned}
\bar{P}(G) &= \frac{m^{2N}}{\binom{|V|}{2}} \left(\frac{N}{(m-1)^2} - \frac{m+1}{(m-1)^3} \right) + o(1) \\
&= \frac{2m^{2N}}{(m^N - 1)^2 - (m^N - 1)(m - 1)} \left(N - \frac{m+1}{m-1} \right) + o(1) \\
&\sim 2 \left(N - 1 - \frac{2}{m-1} \right) \text{ as } N \to \infty,
\end{aligned}$$

hence $\bar{P}(G) \sim D - \frac{4}{m-1}$ as $N \to \infty$. $\square$

Clearly, by substituting $m = 2$ into Theorem 3, we arrive at the asymptotic estimate $\bar{P}(G) \sim D - 4$, when $G = (E, V)$ is a complete binary tree, as first conjectured by T. Lewis [3, p. 83].

References

[1] E. R. Canfield, R. W. Robinson, and D. H. Rouvray, Determination of the Wiener molecular branching index for the general tree, *J. Comput. Chem.* **6** (1985), 598–609.

[2] A. D. Dobrynin, R. Entringer, and I. Gutman, Wiener index of trees: theory and applications, *Acta Applicandae Mathematicae* **66** (2001), 211–249.

[3] T. G. Lewis, *Network Science*, Wiley, 2009.

[4] H. Liu and X. Pan, On the Wiener index of trees with fixed diameter, *MATCH Commun. Math. Comput. Chem.* **60** (2008), 85–94.

2010 *Mathematics Subject Classification*: Primary 05C30; Secondary 05C38.
Keywords: average path length, Wiener index, m-ary tree.

Received January 23 2014; revised version received May 7 2014. Published in *Journal of Integer Sequences*, May 8 2014.

Return to Journal of Integer Sequences home page.