Received: 29 August 2019 | Revised: 8 October 2020 | Accepted: 21 October 2020

DOI: 10.1002/jgt.22643

ARTICLE

WILEY

The average size of a connected vertex set of a graph—Explicit formulas and open problems

Andrew Vince

Department of Mathematics, University of Florida, Gainesville, Florida, USA

Correspondence
Andrew Vince, Department of Mathematics, University of Florida, Gainesville, FL, USA.
Email: avince@ufl.edu

Funding information
Simons Foundation, Grant/Award Number: 322515

Abstract

Although connectivity is a basic concept in graph theory, the enumeration of connected subgraphs of a graph has only recently received attention. The topic of this paper is the average order of a connected induced subgraph of a graph. This generalizes, to graphs in general, the average order of a subtree of a tree. For various infinite families of graphs, we investigate the asymptotic behavior of the proportion of vertices in an induced connected subgraph of average order. For ladders and circular ladders, an explicit closed formula is derived for the average order of a connected induced subgraph in terms of the classic Pell numbers. These formulas imply that, asymptotically, 3/4 of the vertices of a ladder or circular ladder, on average, are present in a connected induced subgraph. Results on such infinite families motivate an assortment of open problems.

KEYWORDS

average order of a connected induced subgraph, connectedness, enumeration, graph

2010 MATHEMATICS SUBJECT CLASSIFICATION

05C30

1 | INTRODUCTION

Although connectivity is a basic concept in graph theory, problems involving the enumeration of connected subgraphs of a graph have only recently received attention. The topic of this paper is the average order of a connected induced subgraph of a graph. Let $G = (V, E)$ be a connected finite simple graph, and $U \subseteq V$ a subset of its vertices. The set U is said to be *connected* if the subgraph of G induced by U is connected. Hereafter, a connected subset of vertices will be referred to simply as a *connected set*. The number of connected sets in G will be denoted by $N(G)$. If $\{U_1, U_2, ..., U_{N(G)}\}$ is the set of all connected sets of a graph G of order n, then we use the notation $S(G) := \sum_{i=1}^{N} |U_i|$. Further, let

$$A(G) = \frac{S(G)}{N(G)} \quad \text{and} \quad D(G) = \frac{A(G)}{n}$$

denote, respectively, the average size of a connected set of G and the proportion of vertices in an average size connected set. The parameter $D(G)$ is referred to as the *density* of connected sets of vertices. The density allows us to compare the average size of connected sets of graphs of different orders. The density is also the probability that a vertex chosen at random from G will belong to a randomly chosen connected set of G. The empty set, in some examples, will be counted as connected, while not in other in other examples. This simplifies calculations and has no effect on results on asymptotic density of families of graphs. By *asymptotic density* of a sequence $\{G_n\}$ of graphs, we mean $\lim_{n\to\infty} D(G_n)$. The following elementary examples illustrate these concepts and are referred to later in the paper.

Example 1 (Paths). If P_n is the path with n vertices, then the number of connected sets is $N(P_n) = \sum_{i=1}^{n} i = n(n+1)/2$, and the sum of the sizes of these connected sets is $S(P_n) = \sum_{i=1}^{n} i(n-i+1) = n(n+1)^2/6$. Therefore

$$A(P_n) = \frac{n+1}{3} \quad \text{and} \quad D(P_n) = \frac{n+1}{3n}.$$

Asymptotically, $D(P_n)$ decreases to $\frac{1}{3}$ as $n \to \infty$. So a connected set of vertices of a path contains approximately a third of the total number of vertices on average.

Example 2 (Cycles and complete graphs). Let C_n denote the n-cycle and K_n the complete graph. Since, for these families, the complement $V \setminus U$ of every connected set U is again connected (taking the empty set to be connected), we have $S(C_n) = (N(C_n)/2)n$ and $S(K_n) = (N(K_n)/2)n$. Therefore

$$D(C_n) = D(K_n) = 1/2.$$

Example 3 (Complete bipartite graphs). It is not hard to show that $N(K_{n,n}) = (2^n - 1)^2 + 2n$ (not counting the empty set as connected), and $S(K_{n,n}) = n(4^n - 2^n + 2)$. Therefore

$$D(K_{n,n}) = \frac{1}{2}\left(\frac{4^n - 2^n + 2}{4^n - 2^{n+1} + 2n + 1}\right). \tag{1}$$

A slightly more complicated formula for $D(K_{m,n})$ can be proved by either (1) splitting the connected sets into those whose complements are connected and those whose complements are not, then arguing as in Example 2, or (2) by induction on the number of vertices. Clearly

$$\lim_{n\to\infty} D(K_{n,n}) = 1/2.$$

There are a number of papers on the average size and density of connected sets of trees. The invariant $A(G)$, in this case, is the average order of a subtree of a tree. We review known results in Section 2. Nearly nothing is known for graphs in general. Our original goal was to investigate the average size of a connected set of the $n \times n$ grid graph $G_n = P_n \times P_n$, where P_n is the path with n vertices and $\times$ denotes the Cartesian product. This is equivalent to finding the average area of a polyomino that fits on an $n \times n$ chess board with unit area squares. Before attempting this, we decided to warm up with a $2 \times n$ board, equivalently the *ladder graph* $L_n = P_n \times P_2$, and also the *circular ladder* $CL_n = C_n \times P_2$ (see Figure 1). Our naive assumption was that this would be closer in difficulty to $P_n = P_n \times P_1$ and $C_n = C_n \times P_1$, which were easily determined in Examples 1 and 2. This turned out not to be the case. These two families are a main topic of this paper. An explicit closed formula is given for $D(L_n)$ and for $D(CL_n)$ in terms of the classical Pell numbers

$$\beta(n) = \frac{1}{2}((1+\sqrt{2})^n + (1-\sqrt{2})^n) \quad \text{and} \quad \overline{\beta}(n) = \frac{1}{2\sqrt{2}}((1+\sqrt{2})^n - (1-\sqrt{2})^n).$$

These numbers have been known since ancient times as the numerators and denominators of the closest rational approximations to $\sqrt{2}$. Our formulas for $D(L_n)$ and for $D(CL_n)$ appear in Theorems 6.4 and 7.5. The asymptotic densities are

$$\lim_{n\to\infty} D(L_n) = \lim_{n\to\infty} D(CL_n) = \frac{3}{4}.$$

The proofs employ generating function methods.

The average size of a connected set and the asymptotic density of the grid graphs remain unsolved. See Section 8 for open problems and conjectures motivated by results in this paper.

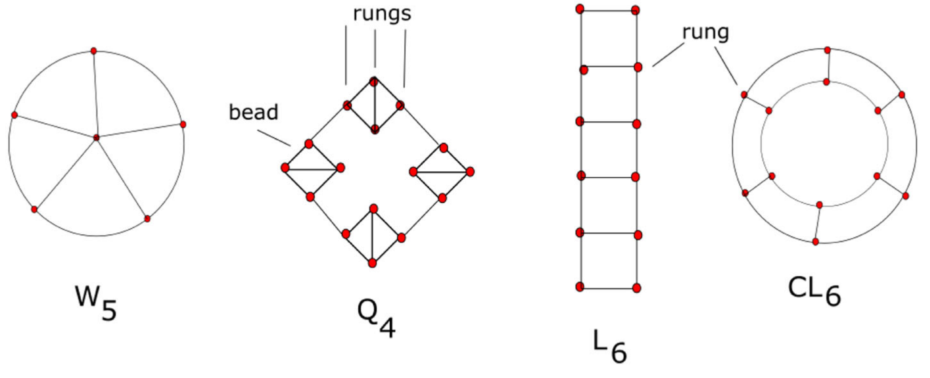


FIGURE 1 Examples of wheels, necklaces, ladders, and circular ladders [Color figure can be viewed at wileyonlinelibrary.com]

After a review of the literature in Section 2, in Section 3 we find the average size and density of connected sets for wheels and necklaces, results relevant to questions in Section 8. Section 4 provides an algorithm for computing the average size of a connected set for the family of ladders. For a reasonably sized n (the answer is almost instantaneous for $n = 500$ on our laptop), the algorithm computes $A(L_n)$ and $D(L_n)$. It does not provide a formula for $A(L_n)$ or $D(L_n)$. The algorithm is symbolic: for a given n, it computes a polynomial $p(x) = \sum_i a_i\, x^i$, where a_i is the number of connected sets of L_n of size i. Therefore the average size of a connected set is

$$A(L_n) = \frac{p'(1)}{2n\; p(1)}.$$

The exact closed formulas and their proofs, for $D(L_n)$ and $D(CL_n)$, are given in Sections 6 and 7. The proofs rely on binomial identities involving the Pell numbers, these preliminary results contained in Section 5. A discussion of conjectures and open problems is the content of Section 8.

2 | PREVIOUS RESULTS

Jamison [2] initiated the study of the average order of a subtree of a tree. A number of papers on the subject followed [1,3,5,6,8,9,11,12]. Concerning lower bounds, Jamison proved that the density, over all trees of order n, is minimized by the path P_n. In particular $D(T) > 1/3$ for all trees T (see Example 1). Concerning upper bounds, he provided a sequence of trees (certain "batons") showing that there are trees with density arbitrarily close to 1. However, if the density $D(T_n)$ of a sequence T_n of trees tends to 1, then the proportion of vertices of degree 2 also has to tend to 1. This led him to conjecture that if T is a tree whose internal vertices have degree at least three, then the average order of a subtree is at least half the order of T. In other words, for such a tree, $D(T) \geq \frac{1}{2}$. Vince and Wang [9] confirmed this conjecture, proving that if T is a tree, all of whose internal vertices have degree at least three, then $\frac{1}{2} \leq D(T) < \frac{3}{4}$. Both bounds are best possible in the sense that there exists an infinite sequence $\{ST_n\}$ of trees (e.g., stars) such that $\lim_{n\to\infty} D(ST_n) = 1/2$ and an infinite sequence $\{CAT_n\}$ of trees (certain "caterpillars") such that $\lim_{n\to\infty} D(CAT_n) = 3/4$. Meir and Moon [5] determined the average density over all trees of order n to be $1 - e^{-1} \approx 0.6321$ as $n \to \infty$.

A subtree of a tree T is a connected induced subgraph of T. So it is natural to extend from trees to graphs G by asking about the average order of a connected induced subgraph of G—or, in our terminology, the average size of a connected set of vertices of G. Kroeker, Mol, and Oellermann [4] initiated such an investigation for *cographs*, that is, graphs that contain no induced P_4. For a connected cograph G of order n, they proved that $n/2 < A(G) \leq (n+1)/2$, with equality on the right if and only if $n = 1$. Complete bipartite graphs (see Example 3) are examples of cographs. In fact, cographs have the following known characterization: a graph G is a cograph if and only if $G = K_1$ or there exist two cographs G_1 and G_2 such that either G is the disjoint union of G_1 and G_2 or G is obtained from the disjoint union by adding all edges joining the vertices of G_1 and G_2. Proving bounds on $A(G)$ for cographs is therefore amenable to an inductive approach not applicable to graphs in general.

3 | TWO MORE EXAMPLES

Two additional illustrative examples are the subject of this section, the family of wheels and the family of necklaces (defined below).

Example 4. Let W_n, $n \geq 3$, denote the wheel with n spokes, as shown in Figure 1. The number of connected sets not containing the hub is $n(n-1)+2 = n^2 - n + 2$ and, by the method of Example 2, the total number of vertices in these sets is $(n^2 - n + 2)(n/2)$. The number of connected sets containing the hub is 2^n, and the total number of vertices in these sets is $2^n + \sum_{i=1}^{n} \binom{n}{i} i = 2^{n-1}(n+2)$. Summing we obtain

$$D(W_n) = \frac{1}{2}\left(\frac{n^3 - n^2 + 2n + (n+2)2^n}{n^3 + n + 2 + (n+1)2^n}\right).$$

For all $n \geq 3$, it is the case that $D(W_n) > \frac{1}{2}$, and the density reaches a maximum at $n = 10$ with $D(W_{10}) \approx 0.538$, that is, the average size of a connected set in this case is slightly more than 1/2. Asymptotically

$$\lim_{n\to\infty} D(W_n) = \frac{1}{2}.$$

Example 5. Let Q_n, $n \geq 1$, be the *necklace* with n *beads*, each bead containing four vertices. The necklace Q_4 in Figure 1 has four beads and 16 vertices. Each bead has three *rungs*, the two end rungs with one vertex each and the center rung with two vertices.

Theorem 3.1. *The density of the necklace Q_n is*

$$D(Q_n) = \frac{150\; n\; 3^{n-1} - 139 \cdot 3^{n-1} + 53}{8\; 3^n + 4n(5 \cdot 3^{n+1} - 21)},$$

and the asymptotic density is

$$\lim_{n\to\infty} D(Q_n) = \frac{5}{6} \approx 0.833333.$$

Proof. Label the $3n$ rungs in Q_n successively around the necklace by $1, 2, 3, \ldots, 3n$, modulo $3n$. Any three consecutive rungs of Q_n can appear as any one of those in Figure 2. For any $i \in \{1, 2, \ldots, 3n\}$ and positive integer $m < n$, let $X(i, m)$ denote the set of rungs $i+1, i+2, \ldots, i+m$. Let $u(m, i)$ denote the number of connected sets with vertices on each rung of $X(i, m)$, and let $w(m, i) = u(m, i) + u(m, i+1) + u(m, i+2)$. It is easy to check that $w(1, i) = 5$, $w(2, i) = 7$, $w(3, i) = 9$, independent of i. Referring to Figure 2,

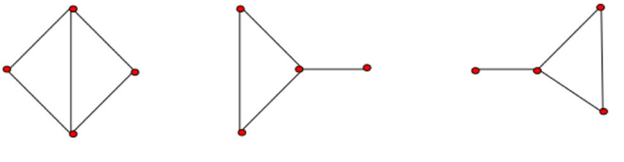


FIGURE 2 Three consecutive rungs as used in the proof of Theorem 3.1 [Color figure can be viewed at wileyonlinelibrary.com]

for any three consecutive rungs, there are exactly three ways to choose a set of vertices containing at least one vertex on each of these rungs. Therefore the recurrence $w(m, i) = 3w(m-3, i)$ holds for $3n > m > 3$ and for all i. Since $w(1, i), w(2, i), w(3, i)$ are independent of i, so is $w(m, i)$ for $3n > m > 3$. So let $u(m) = w(m, i)$, and we have the recurrence

$$u(1) = 5, \quad u(2) = 7, \quad u(3) = 9, \quad u(m) = 3u(m-3) \quad \text{for} \quad 3n > m > 3.$$

This recurrence is routine to solve: $u(3m) = u(3m+1) = 5 \cdot 3^m, u(3m+2) = 7 \cdot 3^m$, $u(3m+3) = 9 \cdot 3^m$, for $1 \leq m < n$. Note that $u(m)$ counts, independent of i, the number of connected sets contained in $Y(i, m) := X(i, m) \cup X(i+1, m) \cup X(i+2, m)$ containing vertices on exactly m consecutive rungs. The following sum, which counts all connected sets not containing vertices on all $3n$ rungs, is also routine to calculate

$$n\sum_{m=1}^{3n-1} u(m) = n\left(\frac{5 \cdot 3^{n+1} - 21}{2}\right).$$

The factor n is required to count all n rotations around the necklace of say $Y(1, m)$.

If a connected set contains vertices on all rungs, then on rungs with two vertices, there are three possibilities. Therefore

$$N(Q_n) = 3^n + n\left(\frac{5 \cdot 3^{n+1} - 21}{2}\right). \tag{2}$$

Let $U(m)$ denote the total number of vertices on all connected sets U with vertices on exactly $m < 3n$ consecutive rungs. The recurrence in this case is $U(m) = 3\, U(m-3) + 10\, u(m-3)$. As above, this can be solved using cases modulo 3, then summed to obtain, after some simplification,

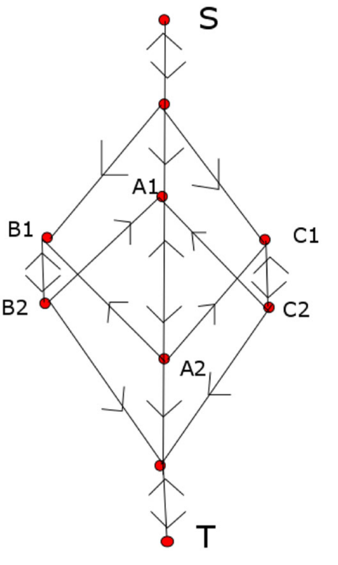
$$\sum_{m=1}^{3n-1} U(m) = 75n \cdot 3^{n-1} - \frac{53}{2}(3^n - 1).$$

Consider next the set of connected sets containing vertices on all rungs. As above, the number of such sets is 3^n. Half of the vertices, the one on rungs with a single vertex ($2n$ such vertices), appear in all such sets. The other half appears in $2 \cdot 3^{n-1}$ such sets. Therefore the total number of vertices in all connected sets is $2n3^n + 2n\ 2 \cdot 3^{n-1} = 10n3^{n-1}$. Therefore

$$S(Q_n) = 10n3^{n-1} + n\left(75n \cdot 3^{n-1} - \frac{53}{2}(3^n - 1)\right) = 75n^2 \cdot 3^{n-1} - \frac{n}{2}(139 \cdot 3^{n-1} - 53). \tag{3}$$

The formula for the density $D(Q_n) = S(Q_n)/(4n\ N(Q_n))$ in the statement of the theorem is routinely obtained from Equations (2) and (3). □

FIGURE 3 Digraph used in the proof of Theorem 4.1
[Color figure can be viewed at wileyonlinelibrary.com]



4 | LADDERS—A GRAPH THEORETIC MATRIX APPROACH

The *ladder* L_n, shown in Figure 1, is the Cartesian product $P_n \times P_2$, consisting of n *rungs*, two vertices per rung. Consider the directed graph G in Figure 3. Arrows in both directions indicate a double edge, one edge in each direction. The matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & x & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & x & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

is the adjacency matrix of G with certain values of 1 replaced by x or x^2, the reason explained in the proof of Theorem 4.1. The vertex S in Figure 3 corresponds to the first row and column of the matrix, and vertex T corresponds to the last row and column. The entry in row r, column c is denoted $A(r, c)$.

Theorem 4.1. *The entry $A^{2n+3}(1, 10)$ is a polynomial $p(x) = \sum_i a_i\, x^i$, where a_i is the number of nonempty connected sets of L_n of size i. In particular,*

$$D(L_n) = p'(1)/(2n\; p(1)).$$

Proof. Let $\mathcal{S}(n)$ denote the set of all sequences of length n in the alphabet $\{0, 1, 1', 2\}$ such that there is no consecutive pair $1, 1'$ or $1', 1$, and 0s can appear only at the beginning and at the end of a sequence. For $s \in \mathcal{S}(n)$, let $|s|$ denote the sum of the terms of s, where both 1 and $1'$ count as 1 and 0 and 2 counts as 0 and 2, respectively. We will refer to $|s|$ as the *weight* of s. Clearly $\mathcal{S}(n)$ is in bijection with the set $\mathcal{C}(n)$ of all connected sets of L_n. Moreover $N(L_n) = |\mathcal{S}(n)|$, and $S(L_n) = \sum_{s\in\mathcal{S}(n)} |s|$.

The digraph in Figure 3 is designed so that each path of length $2n + 3$ from vertices S to T represents a sequence in $\mathcal{S}$. An edge of a path going from vertex $A1$ to $A2$ represents a 2 in the sequence. Similarly an edge from $B1$ to $B2$ represents a 1, and an edge from $C1$ to $C2$ represents a $1'$. An edge going into vertex S represents a 0, as does an edge going out from vertex from vertex T; these can occur only at the beginning and end of the sequence. For example, if S' and T' are the vertices adjacent to S and T, respectively, then the path $S\ S'\ S\ S'\ B1\ B2\ B1\ B2\ A1\ A2\ C1\ C2\ T'\ T$ of length 13 represents the sequence 01121' of length 5. The 3 in $2n + 3$ is due to the initial edge of the path and the last edge of the path. There is a factor 2 in $2n + 3$ because each term in a sequence in $\mathcal{S}$ corresponds to two edges (ignoring the first and last) of the path. Note that, for a path at $A1$ ($B1$, $C1$), the next vertex of the path must be at $A2$ ($B2$, $C2$). If $x = 1$ in the matrix A, then the entry $A^{2n+3}(1, 10)$ is the number of paths of length $2n + 3$ in G, that is, $N(L_n) = p(1)$.

More specifically, a path of length $2n + 3$ from S to T in G corresponds in A^{2n+3} to a monomial of the form $x^e := x^{e_1}\ x^{e_2} \dots x^{e_{2n+3}}$, where each e_i lies in 0, 1, 2. Because the $A1\ A2$ edge is weighted x^2 and the $B1\ B2$ and $C1\ C2$ edges are weighed x, the exponent e is the weight of the corresponding sequence in $\mathcal{S}$. This implies that the coefficient of x^e in $A^{2n+3}(1, 10)$ is the number of sequences in $\mathcal{S}(n)$ with weight e. Hence the derivative $p'(1)$ is the sum of the weights of all the sequences in $\mathcal{S}(n)$. Now $D(L_n) = S(L_n))/(2n\ N(L_n) = p'(1)/(2n\ p(1))$. □

Mathematica produced the following approximate values for the density:

$$
\begin{aligned}
D(L_{100}) &\approx 0.738661,\\
D(L_{200}) &\approx 0.744331,\\
D(L_{300}) &\approx 0.74622,\\
D(L_{400}) &\approx 0.747165,\\
D(L_{500}) &\approx 0.747732.
\end{aligned}
$$

5 | PELL NUMBERS AND BINOMIAL IDENTITIES

Let $\alpha = 1 + \sqrt{2}$ and $\overline{\alpha} = 1 - \sqrt{2}$. Let $\beta(n) = (\alpha^n + \overline{\alpha}^n)/2$ and $\overline{\beta}(n) = (\alpha^n - \overline{\alpha}^n)/2\sqrt{2}$. Note that, while the "overline" in $\overline{\alpha}$ denotes the conjugate in $\mathbb{Z}(\sqrt{2})$, the "overline" in $\overline{\beta}$ is just a notational convention. The sequences $\{\overline{\beta}(n)\}$ and $\{\beta(n)\}$ are Pell numbers and Pell–Lucas numbers, respectively, known since ancient times as the numerators and denominators of the closest rational approximations to $\sqrt{2}$. They appear in many combinatorial settings; see items A000129 and A001333 of OEIS [7] for a long list of references. Both sequences satisfy the same recurrence $\beta(n) = 2\beta(n-1) + \beta(n-2)$ with initial values $\beta(0) = 1, \overline{\beta}(0) = 0$, and $\beta(1) = \overline{\beta}(1) = 1$. The sequences begin $1, 1, 3, 7, 17, 41, \dots$ and $0, 1, 2, 5, 12, 29, \dots$ and have generating functions $(1-x)/(1-2x-x^2)$ and $x/(1-2x-x^2)$, respectively. Using just the formula for the sum of a finite geometric series and its derivatives, it is routine to compute sums of the form $\sum_{m=0}^{n}\beta(m)$, $\sum_{m=1}^{n} m\,\beta(m)$, $\sum_{m=1}^{n} m^2\,\beta(m)$. We omit the exact formulas, but they are used in calculations in the next two sections. It is also easy to verify by induction that

$$\begin{aligned}\beta(n)+\overline{\beta}(n)&=\overline{\beta}(n+1), \quad \beta(n-1)+\beta(n)=2\overline{\beta}(n),\\ \beta(n)-\overline{\beta}(n)&=\overline{\beta}(n-1), \quad \beta(n+1)-\beta(n)=2\overline{\beta}(n).\end{aligned} \tag{4}$$

The primary goal of this section is Theorem 5.8, which is required for the results in Sections 6 and 7. The generating function formulas in the lemmas used to prove this theorem, however, may be of interest in their own right.

Lemma 5.1. *For $n \geq 1$,*

$$\sum_{k\geq 0}\binom{n}{2k}2^k=\beta(n) \quad \text{and} \quad \sum_{k\geq 0}\binom{n}{2k+1}2^k=\overline{\beta}(n).$$

Proof. First note that the values on both sides of the equations agree for $n = 1, 2$. If $E(n)=\sum_{k\geq 0}\binom{n}{2k}2^k$ and $U(n)=\sum_{k\geq 0}\binom{n}{2k+1}2^k$, then $E(n)$ and $U(n)$ satisfy the same recurrence, namely, for $n \geq 3$,

$$\begin{aligned}E(n)&=2E(n-1)+E(n-2),\\ U(n)&=2U(n-1)+U(n-2).\end{aligned}$$

We prove this by induction on n. It is true for $n = 3$; assume the statement true for n. Then $U(n+1)=E(n)+U(n)$ implies that the statement is true for $n+1$ in the case of $U(n)$.

Concerning $E(n)$, note that $E(n)+\sqrt{2}\,U(n)=\sum_{k=1}^{n}\binom{n}{k}\sqrt{2}^k=(1+\sqrt{2})^n$. It is easy to verify that $T(n):=(1+\sqrt{2})^n$ satisfies the same recurrence, namely, $T(n)=2T(n-1)+T(n-2)$. Therefore so does $E(n)=T(n)-\sqrt{2}\,U(n)$. □

Lemma 5.2. *For a given positive integer n we have*

$$\sum_{k\geq 1, j\geq k}\binom{j-1}{k-1}\binom{n-j-1}{k}2^k=\beta(n-1)-1.$$

Proof. The expression $\binom{j-1}{k-1}$ counts the number of compositions of j into $k \geq 1$ parts, no part equal to 0. The expression $\binom{n-j-1}{k}$ counts the number of compositions of the remaining $n-j$ into $k+1$ parts, no part equal to 0. Therefore, the expression on the left counts the sum of the number of compositions of n into an odd number $2k+1$ parts, each summand weighted by 2^k. This, however, is also given by the expression $\sum_{k\geq 1}\binom{n-1}{2k}2^k=\beta(n-1)-1$, the equality from Lemma 5.1. □

For any fixed integer d, let

$$C(m,n,d)=\sum_k \binom{m}{k}\binom{n}{k+d}2^k.$$

The domain of the function $C(m,n,d)$ is the set of triples (m,n,d) of integers such that n is nonnegative and $m \geq \max\{0,-d\}$. The sum is over all integers k, but only finitely many terms are nonzero. In the following lemma and its two corollaries, we consider m and d as fixed and consider the infinite series $C(m,n,d)$, $\overline{C}(m,n)$, and $c(m,n)$ in the variable n. All generating functions in this paper are ordinary generating functions.

Remark 1. The following convention will hold with respect to generating functions. Let $f(n)$ be a function whose domain is a set S of integers; in our case S is of the form $\{n : n \geq n_0\}$ for some nonnegative integer n_0. We say that $F(x)=\sum_{n\geq n_1} a_n x^n$ is a *generating function* of f if $f(n)=a_n$ for all $n \in S$, even if $n_1 < n_0$. For example, according to Lemma 5.3, a generating function of $f(n)=C(3,n,-2)$ is

$$F(x)=\frac{(1+x)^3}{x^2(1-x)^2}=\frac{1}{x^2}+\frac{5}{x}+12+20x+28x^2+36x^3+\cdots.$$

Since we defined $C(3,n,-2)$ only for $n \geq 0$, the two initial terms in the series are "ignored." The analogous convention holds for a two-variable generating function of Lemma 5.6.

Lemma 5.3. *A generating function for* $\{C(m,n,d)\}_{n=0}^{\infty}$ *is*

$$\frac{x^d(1+x)^m}{(1-x)^{m+d+1}}$$

for all integer pairs (d,m) *for which* $m \geq max\{0,-d\}$.

Proof. The proof is by induction on m. If $d \geq 0$, then the initial value in the induction is $m=0$, and it is not hard to check the validity of the lemma for $m=0$ (and for all $d \geq 0$) in this case. If $d<0$, then the first value is $m=-d$, in which case, by Remark 1, it must be verified for all $d<0$ that

$$\frac{x^d(1+x)^{-d}}{1-x}=A+\sum_{n\geq 0}2^{-d}x^n,$$

where A is a sum of a finite number of terms of the form $a_n x^n$, $n<0$. To show this, let $e=-d$ and note that

$$\frac{x^d(1+x)^{-d}}{1-x}=x^{-e}\left(\sum_{i=0}^{e}\binom{e}{i}x^i\right)\sum_{n\geq 0}x^n=A+\sum_{n\geq 0}\left(\sum_{i\geq 0}^{e}\binom{e}{i}\right)x^n=A+\sum_{n\geq 0}2^e x^n.$$

In general, assume that the statement of Lemma 5.3 is true for all integer pairs $(d,m-1)$, where $m-1 \geq \max\{0,-d\}$. Then

$$\sum_k \binom{m}{k}\binom{n}{k+d}2^k = \sum_k \left(\binom{m-1}{k} + \binom{m-1}{k-1}\right)\binom{n}{k+d}2^k$$
$$= \sum_k \binom{m-1}{k}\binom{n}{k+d}2^k + 2\sum_k \binom{m-1}{k}\binom{n}{k+d+1}2^k.$$

Therefore $\sum_{n\ge 0} C(m,n,d)x^n = \sum_{n\ge 0} C(m-1,n,d)x^n + 2\sum_{n\ge 0} C(m-1,n,d+1)x^n$. By the induction hypothesis $\sum_{n\ge 0} C(m-1,n,d)x^n = \frac{x^d(1+x)^{m-1}}{(1-x)^{m+d}}$ and $\sum_{n\ge 0} C(m-1, n, d+1)x^n = \frac{x^d(1+x)^{m-1}}{(1-x)^{m+d+1}}$—with the understanding, via Remark 1, that the equalities above hold for the nonnegative exponent terms in the power series. Therefore

$$\sum_{n\ge 0} C(m,n,d)x^n = \frac{x^d(1+x)^{m-1}}{(1-x)^{m+d}} + 2\frac{x^d(1+x)^{m-1}}{(1-x)^{m+d+1}} = \frac{x^d(1+x)^{m}}{(1-x)^{m+d+1}}. \qquad \square$$

Let

$$\overline{C}(m,n) = C(m,n-1,-1) = \sum_k \binom{m}{k}\binom{n-1}{k-1}2^k.$$

The domain of the function $\overline{C}(m,n)$ is the set of all integer pairs (m,n) such that $m \ge 0$ and $n \ge 1$.

Corollary 5.4. *A generating function for* $\{\overline{C}(m,n)\}_{n=1}^{\infty}$ *is*

$$\left(\frac{1+x}{1-x}\right)^m$$

for all $m \ge 0$.

Proof. Let $d = -1$ in Lemma 5.3. Then $\overline{C}(m,n) = C(m,n-1,-1)$ is the coefficient of x^{n-1} in the power series expansion of $x^{-1}\left(\frac{1+x}{1-x}\right)^m$, equivalently the coefficient of x^n in the power series expansion of $\left(\frac{1+x}{1-x}\right)^m$. Note that, according to Remark 1, only terms with $n \ge 1$ are relevant, that is, the constant term is ignored. $\square$

Let $c(m,n) := \overline{C}(m,n-m)$, that is,

$$c(m,n) = \sum_{k\ge 1} \binom{m}{k}\binom{n-m-1}{k-1}2^k.$$

The domain of the function $c(m,n)$ is all the sets of all integer pairs (m,n) such that $m \ge 0$, $n \ge m+1$. The next corollary follows directly from Corollary 5.4.

Corollary 5.5. *A generating function for* $\{c(m,n)\}_{n=m+1}^{\infty}$ *is*

$$x^m\left(\frac{1+x}{1-x}\right)^m$$

for $m \geq 0$.

Next consider the two-variable generating function

$$f(x,y)=\sum_{m\geq 0, n\geq m+1} c(m,n)x^n y^m.$$

Lemma 5.6. *A two-variable generating function for* $\{c(m,n)\}$ *is*

$$\frac{1-x}{1-x-xy-x^2y}.$$

Proof. Multiply both sides of the equation in Corollary 5.5 by y^m and sum with respect to n. It is a geometric series. □

Lemma 5.7. *A generating function for the sequence* $\left\{\sum_{m=1}^{n-1} n\ c(m,n)\right\}_{n=0}^{\infty}$ *is*

$$\frac{x(1+2x-x^2)}{(1-2x-x^2)^2},$$

and the generating function for $\left\{\sum_{m=1}^{n-1} m\ c(m,n)\right\}_{n=0}^{\infty}$ *is*

$$\frac{x(1-x^2)}{(1-2x-x^2)^2}.$$

Proof. By Lemma 5.6, the term $\sum_{m=1}^{n-1} c(m,n)$ is equal to the coefficient of x^n in the power series expansion of $\frac{1-x}{1-x-xy-x^2y}$ with $y=1$, that is, of $\frac{1-x}{1-2x-x^2}$. Therefore $n\sum_{m=1}^{n-1} c(m,n)$ is equal to the coefficient of x^{n-1} in the power series expansion of the x-derivative of $\frac{1-x}{1-2x-x^2}$. Equivalently, it is equal to the coefficient of x^n in the power series expansion of $\frac{x(1+2x-x^2)}{(1-2x-x^2)^2}$. The term $\sum_{m=1}^{n-1} m\ c(m,n)$ is the coefficient of x^n of the y-derivative of $\frac{1-x}{1-x-xy-x^2y}$ after y is set equal to 1. Equivalently, it is the coefficient of x^n in the power series expansion of $\frac{x(1-x^2)}{(1-2x-x^2)^2}$. □

Theorem 5.8.

$$\sum_{k\geq 1}\sum_{j\geq k} j\binom{j-1}{k-1}\binom{n-j-1}{k}2^k=(n\beta(n-1)-\overline{\beta}(n))/2.$$

Proof. Note that

$$\sum_{j=1}^{n-1} j\left(\sum_{k\geq 1}\binom{j-1}{k-1}\binom{n-j}{k}2^k\right) = \sum_{j=1}^{n-1} jc(n-j,n) = \sum_{j=1}^{n-1}(n-j)c(j,n)$$
$$= n\sum_{j=1}^{n-1} c(j,n) - \sum_{j=1}^{n-1} j\ c(j,n).$$

By Lemma 5.7, $\sum_{j=1}^{n-1} j\left(\sum_{k\geq 1}\binom{j-1}{k-1}\binom{n-j}{k}2^k\right)$ is the coefficient of x^n in the expansion of

$$\frac{x(1+2x-x^2)-x(1-x^2)}{(1-2x-x^2)^2} = \frac{2x^2}{(1-2x-x^2)^2} = \frac{x^2}{8}\left(\frac{2}{(\overline{\alpha}+x)^2} + \frac{2}{(\alpha+x)^2} + \frac{\sqrt{2}}{\alpha+x} - \frac{\sqrt{2}}{\overline{\alpha}+x}\right)$$
$$= \frac{1}{4}\sum_{n=2}^{\infty}(-1)^n\left(\frac{n-1}{\alpha^n} + \frac{n-1}{\overline{\alpha}^n} + \frac{1}{\sqrt{2}\ \alpha^{n-1}} - \frac{1}{\sqrt{2}\ \overline{\alpha}^{n-1}}\right)x^n$$
$$= \frac{1}{2}\sum_{n=2}^{\infty}((n-1)\beta(n) + \overline{\beta}(n-1))x^n$$
$$= \frac{1}{2}\sum_{n=2}^{\infty}((n+1)\beta(n) - \overline{\beta}(n+1))x^n.$$

The second equality is by partial fractions. The third equality is by expanding each term as a geometric series or derivative of a geometric series. The fourth equality is from the definition of β and $\overline{\beta}$. The last equality is from Equation (4). The formula in the statement of the theorem follows by shifting n by 1. □

6 | LADDERS

Lemma 6.1. *The number of nonempty connected sets of the ladder L_n is*

$$N(L_n) = (\beta(n+3) - 4n - 7)/2.$$

Proof. We first count the number of connected sets of L_n with at least one vertex on each rung. Let $a(n)$ denote the number of connected sets with two vertices on the last rung of L_n and $b(n)$ the number of connected sets with exactly 1 vertex on the last rung. Then $a(n+1) = a(n) + b(n)$ and $b(n+1) = 2a(n) + b(n)$ for $n \geq 1$, which implies the second-order recurrence $a(n) = 2a(n-1) + a(n-2)$ for $n \geq 3$. Noting that the same recurrence relation $\beta(n) = 2\beta(n-1) + \beta(n-2)$ for the $\beta(n)$, it follows, after checking $n = 1, 2$, that $a(n) = \beta(n)$. It follows that $b(n) = a(n+1) - a(n)$. The number of connected sets of L_n with at least one vertex on each rung is then $a(n) + b(n) = a(n+1) = \beta(n+1)$.

If $c(k)$ denotes the number of connected sets of L_k with at least one vertex on each rung, then (not counting the empty set)

$$N(L_n) = \sum_{k=1}^{n} c(k)(n-k+1) = \sum_{k=1}^{n} \beta(k+1)(n-k+1) = (\beta(n+3) - 4n - 7)/2.$$

The first equality is by the paragraph above, and the second equality is obtained, after simplification, using the formulas for the sum of a geometric series and its derivative. □

Lemma 6.2. *The sum of the sizes of all the connected sets U of L_n, $n \geq 1$, such that U contains vertices on all rungs of L_n is*

$$(3n\beta(n+1) - \overline{\beta}(n))/2.$$

Proof. In this proof, "connected set" U means that U contains vertices on all rungs of L_n. For each connected set U of vertices, assign a binary sequence $s(U)$ of length n in the alphabet $\{1, 2\}$. A 2 at term i in $s(U)$ designates that both vertices of rung i are in U, and a 1 designates that exactly one vertex of rung i is in U. Consider the maximal subsequences of consecutive 1*s* and the maximal subsequences of consecutive 2*s* in $s(U)$, and denote their successive lengths by $a(U) := (a_1, a_2, \ldots, a_{2k})$ or $a(U) := (a_1, a_2, \ldots, a_{2k+1})$, depending on whether there are an even or odd number of such maximal consecutive subsequences. Note that $a_1 + a_2 + \cdots + a_{2k} = n$ in the even case, and $a_1 + a_2 + \cdots + a_{2k+1} = n$ in the odd case. In other words, $a(U)$ is a composition of n such that each part is at least one. For example, for the sequence 1112212212, the composition of 10 is $(3, 2, 1, 2, 1, 1)$, that is, $3 + 2 + 1 + 2 + 1 + 1 = 10$. Each such composition of n comes from exactly two sequences, one obtained from the other by interchanging 1 and 2. For example, the composition $(3, 2, 1, 2, 1, 1)$ comes from 1112212212 and from 2221121121. Call one such sequence the *complement* of the other. Assume that $s(U)$ and $s(\overline{U})$ are complementary and that $s(U)$ begins with a 1. Then the number of vertices in U is $(a_1 + a_3 + a_5 + \cdots) + 2(a_2 + a_4 + a_6 + \cdots)$ and the number of vertices in $\overline{U}$ is $2(a_1 + a_3 + a_5 + \cdots) + (a_2 + a_4 + a_6 + \cdots)$.

Let s be a binary sequence that begins with a 1. Two cases are considered: when the corresponding composition has an even number of parts, say $2k$, and when the corresponding composition has an odd number of parts, say $2k + 1$. Referring to the ladder in Figure 1, the n vertices on the left side will be called *left* vertices, and those on the right will be called *right* vertices. In the even case there are exactly 2^k connected sets U such that $s(U) = s$, because each of the k maximal subsequences of consecutive 1*s* can all be either on the left or on the right in L_n. Consider the set $\mathcal{U}(k)$ of all sets U and $\overline{U}$ that are assigned the same composition $a(U) = a(\overline{U})$. The sum of the sizes of the vertex sets in $\mathcal{U}(k)$ is

$$\begin{aligned}
&((a_1 + a_3 + a_5 + \cdots + a_{2k-1}) + 2(a_2 + a_4 + a_6 + \cdots + a_{2k}))2^k \\
&\quad + (2(a_1 + a_3 + a_5 + \cdots + a_{2k-1}) + (a_2 + a_4 + a_6 + \cdots + a_{2k}))2^k \\
&\quad = 3n\ 2^k,
\end{aligned}$$

and the sum of the sizes of all connected sets U such that $a(U)$ has an even number of parts is

$$3n\sum 2^k = 3n\sum_{k\geq 1}\binom{n-1}{2k-1}2^k = 6n\sum_{k\geq 0}\binom{n-1}{2k+1}2^k = 6n\overline{\beta}(n-1), \tag{5}$$

where the first sum is taken over all compositions of n with an even number, $2k$, of parts. The last equality is by Lemma 5.1.

Next consider the case of those connected sets U such that $a(U)$ has an odd number, say $2k+1$, of parts. Again assume that U is a connected set of vertices such that $s(U)$ begins with a 1 and that $\mathcal{U}(k)$ is the set of all connected sets U and $\overline{U}$ that are assigned the same composition $a(U) = a(\overline{U})$ with an odd number $2k+1$ of parts. The sum of the sizes of vertex sets in $\mathcal{U}(k)$ is

$$\begin{aligned}
&\sum((a_1 + a_3 + a_5 + \cdots + a_{2k+1}) + 2(a_2 + a_4 + a_6 + \cdots + a_{2k}))2^{k+1}\\
&\quad+\sum(2(a_1 + a_3 + a_5 + \cdots + a_{2k+1}) + (a_2 + a_4 + a_6 + \cdots + a_{2k}))2^k\\
&\quad= 4n\sum 2^k + \sum(a_2 + a_4 + a_6 + \cdots + a_{2k})2^k\\
&\quad= 4n\sum_{k\geq 0}\binom{n-1}{2k}2^k + \sum_{k\geq 1}2^k\sum_{j\geq k} j\binom{j-1}{k-1}\binom{n-j-1}{k}\\
&\quad= 4n\beta(n-1) + \sum_{k\geq 1}2^k\sum_{j\geq k} j\binom{j-1}{k-1}\binom{n-j-1}{k},
\end{aligned} \tag{6}$$

where the first sums are taken over all compositions of n with an odd number, $2k+1$, of parts. The last equality is by Lemma 5.1. The formula for $\sum(a_2 + a_4 + a_6 + \cdots + a_{2k})2^k$ is obtained by noting that the number of times that $a_2 + a_4 + a_6 + \cdots + a_{2k} = j$ occurs equals the number of compositions of j into k parts times the number of compositions of $n-j$ (the odd terms) into $k+1$ parts. Adding the results of Equations (5) and (6), the total sum of the sizes of all connected sets U is

$$\begin{aligned}
&6n\overline{\beta}(n-1) + 4n\beta(n-1) + \sum_{k\geq 1}2^k\sum_{j\geq k} j\binom{j-1}{k-1}\binom{n-j-1}{k}\\
&\quad= 4n\beta(n-1) + 6n\overline{\beta}(n-1) + (n\beta(n-1) - \overline{\beta}(n))/2\\
&\quad= (9n\beta(n-1) + 12n\overline{\beta}(n-1) - \overline{\beta}(n))/2\\
&\quad= (3n\beta(n+1) - \overline{\beta}(n))/2.
\end{aligned}$$

The first equality is by Theorem 5.8 and the last equality is by Equation (4). □

Lemma 6.3. *The sum of the sizes of all the connected sets of* L_n, $n \geq 1$, *is*

$$S(L_n) = ((32 - 45\overline{\beta}(n) - 32\beta(n)) + n(10 + 21\beta(n) + 30\overline{\beta}(n))/4.$$

Proof. If $C(m)$ denotes the sum of the sizes of all the connected sets of L_m with at least one vertex on each rung, then the total number of connected sets of L_n equals

$$\begin{aligned}\sum_{m=1}^{n} C(m)(n-m+1) &= \sum_{m=1}^{n}(n-m+1)(3m\beta(m+1)-\overline{\beta}(m))/2 \\ &= \frac{1}{2}(n+1)\sum_{m=1}^{n}(3m\beta(m+1)-\overline{\beta}(m)) \\ &\quad -\frac{1}{2}\sum_{m=1}^{n}(3m^2\beta(m+1)-m\overline{\beta}(m)) \\ &= (n+1)(3(n+1)\overline{\beta}(n+2)-\frac{3}{2}\beta(n+3)-\frac{1}{2}\beta(n+1)+5)/2 \\ &\quad -(3(n+1)^2\overline{\beta}(n+2)+3(1-2n-2n^2)\overline{\beta}(n+3) \\ &\quad +3n^2\overline{\beta}(n+4)-21)/4 \\ &\quad +((n+1)\beta(n+1)-\overline{\beta}(n+2)+1)/4 \\ &= ((32-45\overline{\beta}(n)-32\beta(n))+n(10+21\beta(n)+30\overline{\beta}(n)))/4.\end{aligned}$$

The first equality is from Lemma 6.2, the third from summing geometric series and their derivatives, and the last from Equation (4). □

Theorem 6.4. *The average size of a connected set of the ladder L_n is*

$$A(L_n) = \frac{(32-45\overline{\beta}(n)-32\beta(n))+n(10+21\beta(n)+30\overline{\beta}(n))}{2(\beta(n+3)-4n-7)},$$

and the asymptotic density is

$$\lim_{n\to\infty} D(L_n) = \frac{3}{4}.$$

Proof. The first equality follows directly from Lemmas 6.1 and 6.3. Using the fact that $\lim_{n\to\infty}\overline{\alpha}^n = 0$ and dividing numerator and denominator of $A(L_n)/2n$ by $n\alpha^n$ gives

$$\lim_{n\to\infty} D(L_n) = \frac{30+21\sqrt{2}}{4\sqrt{2}\ \alpha^3} = \frac{3}{4}.$$

□

7 | CIRCULAR LADDERS

A *rung* of a circular ladder is a set of two radial vertices as depicted in Figure 1. The vertices on the inner cycle will be called *inner* vertices, and those on the outer cycle will be called *outer* vertices.

Lemma 7.1. *The number of connected vertex sets U of $CL(n)$ with at least one vertex of U on each rung is* $2\beta(n)+1+n(\beta(n-1)-1)$.

Proof. Let U be a connected set of vertices with at least one vertex on each rung. One of the following two cases must hold: (1) every two consecutive rungs contain two vertices of U that are adjacent, or (2) there exist two consecutive rungs containing exactly two nonadjacent vertices of U. Moreover, since U is connected, there is exactly one pair of consecutive rungs such that case (2) can occur. Call such a consecutive pair a *singular pair*, and the term *singularity* refers to the position between the two rungs of a singular pair.

Consider two cases, the first being where there is no singular pair. In that case, there are $\binom{n}{2k}$ ways to partition the set of rungs of the circular ladder into an even number, say $2k$, of nonempty parts, each part consisting of consecutive rungs. (Just choose k "cuts" between consecutive rungs). Associate to each such partition, the two connected sets U and its complement $\overline{U}$ as in the proof of Lemma 6.2. In this case the number of connected sets with at least one vertex on each rung of $CL(n)$ is

$$2\sum_{k\geq 1}\binom{n}{2k}2^k + 3 = 2\sum_{k\geq 0}\binom{n}{2k}2^k + 1 = 2\beta(n) + 1.$$

The factor 2 is to count both U and its complement $\overline{U}$. The binomial coefficient counts the number of ways to partition the rungs of $CL(n)$ into exactly $2k$ nonempty parts. The 3 is added to count the connected set that has two vertices on every rung and the two connected sets that have exactly one vertex on each rung.

Next consider the second case, where U has a singularity. In this case, partition the rungs of the circular ladder into an odd number, say $2k + 1$, nonempty parts, each part consisting of consecutive rungs. Arranging the parts linearly, we may assume that the singularity is between the first and last rungs. In this case, the number of connected sets with at least one vertex on each rung of $CL(n)$ is

$$n\sum_{k\geq 1}\binom{n-1}{2k+1-1}2^k = n\left(\sum_{k\geq 0}\binom{n-1}{2k}2^k - 1\right) = n(\beta(n-1) - 1).$$

The standard formula $\binom{n-1}{r-1}$ for the number of compositions of n into r, no part 0, is used. The factor n appears because the singularity can occur between any two rungs of $CL(n)$. □

Lemma 7.2. *The number of connected sets of $CL(n)$ is*

$$N(CL_n) = 1 - 3n + 2\beta(n) + 3n\,\overline{\beta}\,(n).$$

Proof. It is part of the proof of Lemma 6.1 that the number of connected sets of L_k with at least one vertex on each rung is $\beta(k+1)$. Therefore, the number of connected sets U of $CL(n)$ with at least one rung containing no vertices of U is

$$n\sum_{k=1}^{n-1}\beta(k+1) = n\left(\sum_{k=0}^{n}\beta(k) - 2\right) = n(\overline{\beta}\,(n+1) - 2) = n\overline{\beta}\,(n+1) - 2n.$$

The second equality is by summing geometric series.

Combining the above equation with Lemma 7.1, the number of connected vertex sets of $CL(n)$ is

$$\begin{aligned} 2\beta(n) + 1 + n(\beta(n-1) - 1) - 2n + n\overline{\beta}(n+1) &= 1 - 3n + 2\beta(n) + n\beta(n-1) + n\overline{\beta}(n+1) \\ &= 1 - 3n + 2\beta(n) + 3n\overline{\beta}(n). \end{aligned}$$ □

Lemma 7.3. *The sum of the sizes of the connected sets U of $CL(n)$ such that U contains at least one vertex on each rung is*

$$(2n + 6n\beta(n) + 3n^2\beta(n-1) - 2n^2 - n\overline{\beta}(n))/2.$$

Proof. The terminology used in the proof of Lemma 7.1 is also used in this proof. Consider the two cases: where there is no singularity and where there is. If there is no singularity, then the sum of the sizes of the connected sets U of $CL(n)$ is

$$4n + 3n\sum_{k\geq 1}\binom{n}{2k}2^k = 4n + 3n(\beta(n) - 1) = n(1 + 3\beta(n)). \tag{7}$$

The summation is as in the proof of Lemma 7.1; the factor $3n$ comes from that fact that the number of vertices in U and its complement $\overline{U}$ for any particular subdivision of the rungs of $CL(n)$ is $3n$. The term $4n$ is added for the U with two vertices in each rung of $CL(n)$ and for the U with exactly one vertex in each rung.

In the case where U has a singular point, label the rungs $\{1, 2, \ldots, n\}$ and assume that there are exactly j rungs having two vertices in U. Partition $\{1, 2, \ldots, n\}$ into exactly $2k + 1$ consecutive nonempty parts, the first and last parts corresponding to rungs containing exactly one vertex of U. The singularity occurs between the first and last rungs. This can be done by first partitioning the j rungs containing two vertices of U into k parts, then partitioning the remaining $n - j$ rungs into $k + 1$ parts. Therefore the sum of the number of vertices over all such subsets U is

$$\begin{aligned} n\left(\sum_{k\geq 1}2^k\sum_{j\geq k}(n+j)\binom{j-1}{k-1}\binom{n-j-1}{k}\right) &= n^2\left(\sum_{k\geq 1}2^k\sum_{j\geq k}\binom{j-1}{k-1}\binom{n-j-1}{k}\right) \\ &\quad + n\left(\sum_{k\geq 1}2^k\sum_{j\geq k}j\binom{j-1}{k-1}\binom{n-j-1}{k}\right) \\ &= n^2(\beta(n-1) - 1) + n(n\beta(n-1) - \overline{\beta}(n))/2 \\ &= (3n^2\beta(n-1) - 2n^2 - n\overline{\beta}(n))/2. \end{aligned} \tag{8}$$

The factor n at the beginning of the formula comes from the fact that the singularity can occur between any pair of consecutive rungs. The factor 2^k comes from the fact that each maximal consecutive string of vertices with one vertex on each rung can be reversed, inner interchanged with outer. The factor $n + j = 2j + (n - j)$ counts the number of vertices of each such subset U. The last inequality holds because the first summand in the expression before the equal sign counts the number of connected sets of $CL(n)$ with at least one vertex on each rung and with exactly one singular point. To see this, note that $\sum_{j=k}^{n-k-1}(n+j)\binom{j-1}{k-1}\binom{n-j-1}{k}$ counts the number of ways to partition $\{1, 2, \ldots, n\}$ into an odd number of consecutive intervals. Consider these to represent the rungs of $CL(n)$ with the singular point between 1 and n (considered modulo n). The odd intervals correspond to rungs containing exactly one vertex of our set U. There are two possibilities for each such interval except the first and last. For the pair consisting of the first and last, there are also two possibilities because of the singular point. This justifies the factor 2^k. The second equality comes from Lemma 5.2 and Theorem 5.8.

The formula in the statement of the lemma is obtained by adding the expressions (7) and (8). □

Lemma 7.4. *The sum of the sizes of the connected vertex sets of $CL(n)$ is*

$$S(CL_n) = (n(7 + \beta(n) - 7\overline{\beta}(n)) + n^2(-2 + 9\overline{\beta}(n)))/2.$$

Proof. If $C(m)$ denotes the sum of the sizes of all the connected sets of L_m with at least one vertex on each rung, then by Lemmas 6.2 and 7.3 the sum of the sizes of all connected sets of CL_n equals

$$\begin{aligned}
S(CL_n) &= (2n + 6n\beta(n) + 3n^2\beta(n-1) - 2n^2 - n\overline{\beta}(n))/2 + n\sum_{m=1}^{n-1} C(m)\\
&= (2n + 6n\beta(n) + 3n^2\beta(n-1) - 2n^2 - n\overline{\beta}(n))/2\\
&\quad + \frac{n}{2}\sum_{m=1}^{n-1}(3m\beta(m+1) - \overline{\beta}(m))\\
&= (2n + 6n\beta(n) + 3n^2\beta(n-1) - 2n^2 - n\overline{\beta}(n))/2\\
&\quad + \frac{n}{4}\sum_{m=1}^{n-1}(3(m\alpha^{m+1} + m\overline{\alpha}^{m+1}) - (\alpha^m - \overline{\alpha}^m))\\
&= (2n + 6n\beta(n) + 3n^2\beta(n-1) - 2n^2 - n\overline{\beta}(n))/2\\
&\quad + \frac{n}{4}\sum_{m=1}^{n-1}(3\alpha^2(m\alpha^{m-1} + m\overline{\alpha}^2\overline{\alpha}^{m-1}) - (\alpha^m - \overline{\alpha}^m))\\
&= (2n + 6n\beta(n) + 3n^2\beta(n-1) - 2n^2 - n\overline{\beta}(n))/2\\
&\quad + \frac{n}{4}((3(n-1)\beta(n+2) - 3n\beta(n+1) + 3) + (1 - \beta(n)))\\
&= (2n + 6n\beta(n) + 3n^2\beta(n-1) - 2n^2 - n\overline{\beta}(n))/2\\
&\quad + \left(3n^2\overline{\beta}(n+1) - \frac{3n}{2}\beta(n+2) - \frac{1}{2}n\beta(n) + 5n\right)/2\\
&= (n(7 + \beta(n) - 7\overline{\beta}(n)) + n^2(-2 + 9\overline{\beta}(n)))/2.
\end{aligned}$$

The first equality is by Lemma 7.3, the second equality from Lemma 6.2, the fifth equality by summing geometric series and its derivative, and the last two equalities by Equation (4). □

Theorem 7.5. *The average size of a connected vertex set of CL_n is*

$$A(CL_n) = \frac{n(7 + \beta(n) - 7\overline{\beta}(n)) + n^2(-2 + 9\overline{\beta}(n))}{2(1 - 3n + 2\beta(n) + 3n\overline{\beta}(n))}.$$

The asymptotic density of the family of circular disks is

$$\lim_{n\to\infty} D(CL_n) = \frac{3}{4}.$$

Proof. The first equality is from Lemmas 7.2 and 7.4. Using the fact that $\lim_{n\to\infty} \bar{\alpha}^n = 0$ and dividing numerator and denominator of $A(CL_n)/2n$ by $n^2\alpha^n$ gives

$$\lim_{n\to\infty} D(CL_n) = \frac{9}{12} = \frac{3}{4}.$$ □

8 | OPEN PROBLEMS

The density of the grid graphs remains open.

Question 1. What is the asymptotic density

$$\lim_{n\to\infty} D(P_n \times P_n)?$$

Given the difficulty of finding an explicit formula for $D(P_n \times P_2)$, a question is whether it is possible to determine the asymptotic density of the grid graph $P_n \times P_n$ *without* finding an explicit formula. Perhaps the generating function can be determined and techniques of analytic combinatorics applied.

Even finding the number of connected sets of the $n \times n$ grid is problematic. See [10] for related matters.

Question 2. Find a formula for the number $N(P_n \times P_n)$ of connected sets of the $n \times n$ grid.

Question 3. Find a formula for the number of connected sets $N(P_2 \times P_2 \times \cdots \times P_2)$ of the n-cube.

Question 4. For a given simple, connected graph of order n, let a_k, $k = 0, 1, 2, \ldots, n$, denote the number of connected sets of size k. Is the sequence $a_0, a_1, \ldots, a_n$ necessarily unimodal?

For a given graph G, let $c(G) := \sqrt[n]{N(G)}/2$. The value of $c(G)$ is at most 1. It is a measure of the connectedness of G, the closer to 1, the more connected. In this sense, what, for example, are the most connected 3-regular graphs? More precisely:

Question 5. Let $\mathcal{G}$ denote the set of all simple, connected 3-regular graphs. Find $c_3 := \limsup_{G\in\mathcal{G}} c(G)$.

Lemma 7.2 for circular ladders shows that $c_3 > 0.777$. Which families imply a larger value for c_3? Do expander graphs play a role?

The family of paths has asymptotic density 1/3 and, as discussed in Section 2, it is known that, among all trees of order n, the path minimizes the density. The following was conjectured in [4].

Conjecture 6. *Among all graphs of order* n, *the path minimizes the density.*

If all internal vertices of a tree T have degree at least three, then, as discussed in Section 2, the density of T is at least 1/2. In this paper, it was shown that the families of cycles, complete graphs, complete bipartite graphs $K_{n,n}$, and wheels all have asymptotic density 1/2; the family of ladders and circular ladders has asymptotic density 3/4; and the family of necklaces has asymptotic density 5/6.

Conjecture 7. *If all vertices of a graph* G *have degree at least* 3, *then* $D(G) \geq 1/2$.

Concerning upper bounds, it is natural to ask the following.

Question 8. Do there exist graphs with minimum degree at least 3 and density arbitrarily close to 1?

For a graph G of order n, call a connected set U of vertices *simply connected* if its complement $\overline{U} = V(G)\backslash U$ is also connected. Let $N_1 = N_1(G)$ denote the number of simply connected sets in G; also let $N_2 = N_2(G) = N(G) - N_1(G)$. Given a connected graph G, N_2 is the number of connected vertex sets U such that the subgraph of G induced by $V\backslash U$ is not connected. Call such a U a *separating set*. Let $S_1 = S_1(G)$ denote the sum of the sizes of all simply connected sets and let $S_2 = S - S_1$ be the sum of the sizes of all the separating sets. Note that $S_1 = n\ N_1/2$ because, for a simply connected set U, we have $|U| + |\overline{U}| = n$. Let $D_2 = S_2/(nN_2)$ be the density of the separating sets. Now

$$D(G) = \frac{\frac{1}{2}N_1 + D_2 N_2}{N},$$

which implies

$$1 - D(G) = (1 - D_2)\frac{N_2}{N} + \frac{1}{2}\left(1 - \frac{N_2}{N}\right).$$

Therefore, if there do exist graphs with minimum degree at least 3 and density arbitrarily close to 1, then there must exist graphs for which, simultaneously, the separating set density D_2 and the separating set proportion N_2/N are both arbitrarily close to 1. In other words, there must be very many, very large, separating sets. The necklaces Q_n have a reasonable number of large separating sets, and the density is 5/6. By how much can this be increased?

ACKNOWLEDGMENTS

This study was partially supported by a grant from the Simons Foundation (322515 to Andrew Vince). We thank the referees for their careful reading and their valuable corrections and suggestions.

ORCID

Andrew Vince http://orcid.org/0000-0002-1022-1320

REFERENCES

1. J. Haslegrave, *Extremal results on average subtree density of series-reduced trees*, J. Combin. Theory Ser. B **107** (2014), 26–41.
2. R. Jamison, *On the average number of nodes in a subtree of a tree*, J. Combin. Theory Ser. B **35** (1983), 207–223.
3. R. Jamison, *Monotonicity of the mean order of subtrees*, J. Combin. Theory Ser. B **37** (1984), 70–78.
4. M. E. Kroeker, L. Mol, and O. Oellermann, *On the mean connected induced subgraph order of cographs*, Australas. J. Combin. **71** (2018), 161–183.
5. A. Meir and J. W. Moon, *On subtrees of certain families of rooted trees*, Ars Combin. **16** (1983), 305–318.
6. L. Mol and O. Oellermann, *Maximizing the mean subtree order*, J. Graph Theory (2018). https://doi.org/10.1002/jgt.22434
7. N. J. A. Sloane, *The on-line encyclopedia of integer sequences*, (2020). http://oeis.org
8. A. M. Stephens and O. Oellermann, *The mean order of sub-k-trees of k-trees*, J. Graph Theory **88** (2018), 61–79.
9. A. Vince and H. Wang, *The average order of a subtree of a tree*, J. Combin. Theory Ser. B **100** (2010), 161–170.
10. A. Vince, *Counting connected sets and connected partitions of a graph*, Australas. J. Combin. **67** (2017), 281–293.
11. S. Wagner and H. Wang, *On the local and global means of subtree orders*, J. Graph Theory **81** (2016), 154–166.
12. W. Yan and Y. Yeh, *Enumeration of subtrees of trees*, Theoret. Comput. Sci. **369** (2006), 256–268.

How to cite this article: Vince A. The average size of a connected vertex set of a graph—Explicit formulas and open problems. *J Graph Theory*. 2020;1–22. https://doi.org/10.1002/jgt.22643