The Number of Ways to Assemble a Graph

Andrew Vince and Miklós Bóna
University of Florida, Department of Mathematics
Gainesville, FL, USA
avince@ufl.edu, bona@ufl.edu

Abstract

Motivated by the question of how macromolecules assemble, the notion of an *assembly tree* of a graph is introduced. Given a graph G, the paper is concerned with enumerating the number of assembly trees of G, a problem that applies to the macromolecular assembly problem. Explicit formulas or generating functions are provided for the number of assembly trees of several families of graphs, in particular for what we call (H, ϕ)-graphs. In some natural special cases, we use a powerful recent result of Doron Zeilberger and Moa Apagodu to provide recurrence relations for the *diagonal* of the relevant multivariate generating functions, and we use a result of Jet Wimp and Zeilberger to find very precise asymptotic formulae for the coefficients of these diagonals. Future directions for reseach, as well as open questions, are suggested.

1 Introduction

Although the context of this paper is graph theory, the concept of an assembly tree originated in an attempt to understand macromolecular assembly [4]. The capsid of a virus - the shell that protects the genomic material - self-assembles spontaneously, rapidly and quite accurately in the host cell. Although the structure of the capsid is fairly well known, the assembly process by which hundreds of subunits (monomers) interact to form the capsid is not well understood. In many cases, the capsid can be modeled by a polyhedron, the facets representing the monomers. The assembly of the capsid can be modeled by a rooted tree, the leaves representing the facets, the root the completed polyhedron, and the internal nodes intermediate subassemblies. The enumeration of such trees plays a central role in understanding how symmetry effects the assembly process [4].

All graphs in this paper are simple. Let $G = (V, E)$ be a connected graph of order n with vertex set V and edge set E. In the definition of an assembly tree T for the graph G, each node of T is labeled by a subset of V. No distinction will be made between the node and its label. For a node U in a rooted tree, $c(U)$ denotes the set of children of U.

Definition 1. An *assembly tree* for a connected graph G on n vertices is a rooted tree, each node of which is labeled by a subset $U \subset V$ such that

1. each internal (non-leaf) node has at least two children,

2. there are n leaves with labels $\{v\}$, $v \in V$,

3. the label on the root is V,

4. $U = \bigcup c(U)$ for for each internal node U.

An assembly tree T for G describes a process by which G assembles. At the beginning are the individual vertices of G - the leaves of T. Each internal node U of T represents the subgraph of G induced by the subset U of vertices. Each internal node also represents the stage in the assembly process by which subgraphs of G join to form a larger subgraph; more precisely, the subgraphs induced by the children of U join to form the subgraph induced by U. The process terminates at the root of T - representing the entire graph G. Call two assembly trees T_1 and T_2 for a graph G *equal* if there is a label preserving graph isomorphism between T_1 and T_2.

There are numerous ways, some mentioned in the last section, to further restrict how the assembly process occurs. In this paper we will assume that, at each stage, two subgraphs can be joined if and only if there is an edge that connects them.

Definition 2. An assembly tree for a connected graph $G = (V, E)$ using the *edge gluing rule* is an assembly tree for G satisfying the additional property:

5. Each internal node has exactly two children, and if U_1 and U_2 are the children of internal node U, then there is an edge $\{v_1, v_2\} \in E$, the *gluing edge*, such that $v_1 \in U_1$ and $v_2 \in U_2$.

Figure 1 shows a graph G and two assembly trees for G using the edge gluing rule. Throughout this paper, until the last section, all assembly trees use the edge gluling rule. Therefore, the term "assembly tree" will refer to an assembly tree using the edge gluing rule.

The subject of the paper is, given a graph G, to enumerate the number

$$a(G)$$

of assembly trees of G. Gluing sequences are defined in Section 2 and are used too enumerate the number of assembly trees for paths, cycles and certain star graphs. The concept of an H-graph is defined in Section 3; the complete multi-partite graphs are special cases. A generating function formula for the number of assembly trees for any H-graph is provided in Section 3. Section **??** considers three specific examples for H-graphs which lead to frequently encountered families of graphs, such as complete bi-partite graphs or complete tripartite graphs. For each of these examples, the relevant multivariate generating function is computed, then the *diagonal* of that generating function is introduced and studied. Very strong recent results of Doron Zeilberger and Moa Apagodu enable us to prove polynomial recurrence relations for the coefficients of these diagonals, while results of Zeilberger and Jet Wimp allow us find the growth rate of these coefficients at an arbitrary level of precision. Open questions and further research directions are offered in Section 5.

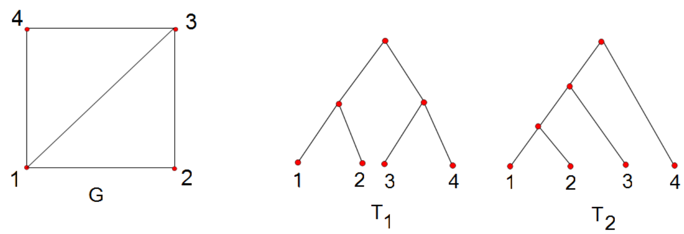

Figure 1: Two assembly trees for the graph G.

2 Paths, Cycles and Stars

An assembly tree for a graph $G = (V, E)$ of order n, as defined in the introduction, is a binary tree with n leaves and $n-1$ internal nodes. To each internal node U there is a corresponding gluing edge as in Definition 2 (not necessarily unique), which we denote by $e_U \in E$.

Lemma 1. *If T is an assembly tree for a connected graph $G = (V, E)$, then the set of gluing edges $\{e_U \,|\, U$ is an internal node of assembly tree $T\}$ is a spanning tree of G.*

Proof. If the set $S := \{e_U \,|\, U$ is an internal node of assembly tree $T\}$ of gluing edges is not spanning, then the root of T would not be V. If S contains a cycle, then T would have a node with just one child. If S is not connected, then G would not be connected. □

If $S \subseteq E(G)$ is any spanning tree of a connected graph G, then any linear ordering $e_1, e_2, \ldots, e_{n-1}$ of the edges in S induces an assembly tree for G as follows. Build the tree T from the bottom up. The leaves are the singleton vertices of G. Assume that we have proceeded through the sequences of edges from e_1 to e_{k-1}. For $e_k = \{u_1, u_2\}$ add a node to T whose two children are the already constructed nodes U_1 and U_2 such that $u_1 \in U_1$ and $u_2 \in U_2$. Call an ordering $e_1, e_2, \ldots, e_{n-1}$ of the edges of the spanning tree a *gluing sequence*. The elements of a gluing sequence for G are the gluing edges of the corresponding assembly tree T.

Example 3. Consider the 4-cycle C_4 in Figure 2. This example shows that two different spanning trees can induce the same assembly tree: the gluing sequences (e_1, e_2, e_3) and

(e_1, e_2, e_4) produce the same assembly tree. Moreover, two different orderings, for example (e_1, e_2, e_3) and (e_2, e_1, e_3), of the same spanning tree can produce the same assembly tree.

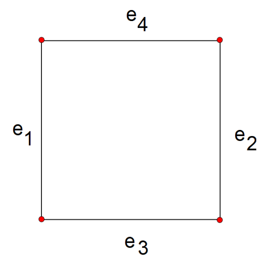

Figure 2: The cycle C_4.

For the star $S_n := K_{1,n}$, the spanning tree is S_n itself, and each gluing sequence produces a distinct assembly tree. This leads immediately to the following result.

Proposition 1. *For the star the number of assembly trees is* $a(S_n) = n!$.

Consider the star S_n^2 with n arms such that each arm has length 2, as in Figure 3.

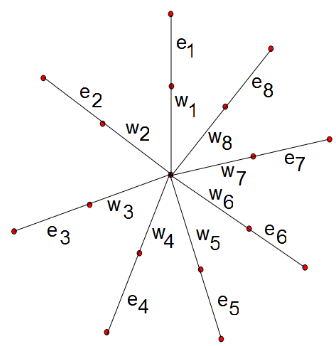

Figure 3: The star S_8^2.

Theorem 4.

$$a(S_n^2) = \sum_{k=0}^{n} \binom{n}{k} \frac{(2n-k)!}{2^{n-k}}$$

Proof. Suppose that k of the e-edges come first in the gluing sequence. In Figure 2 we refer to the e-edges and the w-edges. There are $\binom{n}{k}$ ways to choose these edges, and the order does not matter for the assembly tree. There are $2n-k$ edges that remain. For convenience label them $e_1, e_2, \dots, e_{n-k}$ and $w_1, w_2, \dots, w_n$. These $2n-k$ edges can be placed in any order in the gluing sequence as long as the edge w_i comes after e_i for $i = 1, 2, \dots, n-k$. Each such order determines a distinct assembly tree. To determine the number of such permutations, first choose k positions for $w_{n-k+1}, \dots, w_n$. There are $\binom{2n-k}{k}$ ways to do this, and for each such choice the edges $w_{n-k+1}, \ldots, w_n$ can be permuted in $k!$ ways. The remaining $2(n-k)$ positions are to be filled by the edges $e_1, e_2, \ldots, e_{n-k}$ and $w_1, w_2, \ldots, w_n$ so that the edge w_i comes after e_i for $i = 1, 2, \ldots, n-k$. The number of ways to do this equals the number of permutations of $2(n-k)$ objects where there are 2 objects of type 1, 2 objects of type 2 . . . , 2 objects of type $n-k$. This is equal to $\frac{(2n-2k)!}{2^{n-k}}$. So, with k of the e-edges coming first in the gluing sequence there are

$$\binom{n}{k}\binom{2n-k}{k}k!\frac{(2n-k)!}{2^{n-k}} = \binom{n}{k}\frac{(2n-k)!}{2^{n-k}}$$

assembly trees. Summing over all possible values of k from $k = 0$ to $k = n$ gives the formula in the statement of the theorem. □

Theorem 5. *If P_n is the path and C_n is the cycle on n vertices, then*

$$a(P_n) = \frac{1}{n}\binom{2n-2}{n-1}, \qquad a(C_n) = \frac{1}{2}\binom{2n-2}{n-1}.$$

Proof. For the path, the unique spanning tree S consists of all the edges of the path. We proceed by induction. First consider the number of assembly trees in the case that $e \in S$ is the last edge in the gluing sequence. If the removal of e from G results in subgraphs of orders k and $n-k$, then the number of assembly trees such that e is the last edge in the gluing sequence is $a(P_k)\,a(P_{n-k})$. If T and T' are assembly trees coming from gluing sequences with distinct last elements, then $T \neq T'$. Therefore $a(P_n) = \sum_{k=1}^{n-1} a(P_k)\,a(P_{n-k})$, which is a well know recurrence for the Catalan numbers.

Concerning the cycle, there are n spanning trees of C_n. Given any one of these spanning trees, by the result above for the path, there are $\frac{1}{n}\binom{2n-2}{n-1}$ corresponding assembly trees for C_n, hence a total of $n\,\frac{1}{n}\binom{2n-2}{n-1} = \binom{2n-2}{n-1}$ assembly trees. But each of these is counted twice for following reason. The assembly tree corresponding to a sequence of edges in a spanning tree for which e is the last edge in the gluing sequence and f is the edge of G not in the gluing sequence is equal to the assembly tree for which f is the last edge in the gluing sequence and e is the edge of G not in the gluing sequence. □

3 H-graphs

Let H be a connected graph with vertex set $[N] := \{1, 2, \ldots, N\}$, and let $\phi : [N] \to \{0, 1\}$ be a labeling of the vertices of H. For any sequence $(n_1, n_2, \ldots, n_N)$ of non-negative integers, define a graph $G_{(H,\phi)}(n_1, n_2, \ldots, n_N)$ as follows. The vertex set is

$$V(G_{(H,\phi)}(n_1, n_2, \ldots, n_N)) := \{(i, j) : i \in [N],\, 1 \leq j \leq n_i\}$$

and (i, j) is adjacent to (i', j') if and only if

$$\begin{aligned} i = i' \quad &\text{and} \quad \phi(i) = 1, \quad \text{or} \\ i \neq i' \quad &\text{and} \quad \{i, i'\} \in E(H). \end{aligned}$$

This is equivalent to saying that the graph $G_{(H,\phi)}(n_1, n_2, \ldots, n_N)$ is obtained by replacing each vertex i of H by a complete graph of order n_i or its complement (n_i isolated vertices), and by replacing each edge of H by all possible edges between the graphs that replace the two end vertices of that edge. Call a graph an $(\mathbf{H}, \phi)-$*graph* if it is of the form $G_{(H,\phi)}(n_1, n_2, \ldots, n_N)$ for some choice of the parameters $\{n_1, \ldots, n_N\}$.

The following notation will be used in this section

1. $\mathbf{n} := (n_1, n_2, \ldots, n_N)$
2. $\mathbf{n} \geq \mathbf{k}$ if and only if $n_i \geq k_i$ for all i
3. $\mathbf{x} := (x_1, x_2, \ldots, x_N)$
4. $\mathbf{n}! = n_1! n_2! \cdots n_N!$
5. $\binom{\mathbf{n}}{\mathbf{k}} = \binom{n_1}{k_1}\binom{n_2}{k_2}\cdots\binom{n_N}{k_N}$
6. $\mathbf{x}^{\mathbf{n}} = x_1^{n_1} x_2^{n_2} \cdots x_N^{N^n}$
7. $a_{(H,\phi)}(\mathbf{n}) = a(G_{(H,\phi)}(n_1, n_2, \ldots, n_N))$
8. $A_{(H,\phi)}(\mathbf{x}) = \sum_{\mathbf{n}\geq\mathbf{0}} a_{(H,\phi)}(\mathbf{n}) \, \frac{\mathbf{x}^{\mathbf{n}}}{\mathbf{n}!}$.

The last entry in the above list is the exponential generating function for the number of assembly trees of an (H, ϕ)-graph. The zero vector is denoted $\mathbf{0}$.

Theorem 6. *The exponential generating function for a connected (H, ϕ)-graph is*

$$A_{(H,\phi)}(\mathbf{x}) = 1 - \sqrt{1 - 2\sum_{i=1}^{N} x_i + \sum_{\phi(i)=0} x_i^2 + 2 \sum_{\{i,j\}\notin E(H)} x_i x_j}.$$

Proof. Let $\mathbf{0}$ be the all zeros vector; let $\mathbf{e}_i$ be the vector with each coordinate 0 except the i^{th} coordinate 1; and let $\mathbf{e}_{i,j}$ be the vector with each coordinate 0 except the i^{th} and j^{th} coordinate 1. Note that

$$\begin{aligned}
a_{(H,\phi)}(\mathbf{0}) &= 0 \\
a_{(H,\phi)}(\mathbf{e}_i) &= 1 \quad \text{for all } i \\
a_{(H,\phi)}(2\,\mathbf{e}_i) &= 0 \quad \text{if } \phi(i) = 0 \\
a_{(H,\phi)}(\mathbf{e}_{i,j}) &= \begin{cases} 1 & \text{if } \{i,j\} \in E(H) \\ 0 & \text{if } \{i,j\} \notin E(H) \end{cases}
\end{aligned}$$

The following recurrence holds for all $\mathbf{n}$ except those of the form $\mathbf{e}_i$, $\mathbf{e}_{i,j}$ when $\{i, j\} \notin E(H)$ and $2\mathbf{e}_i$ when $\phi(i) = 0$. To simplify notation, denote $a_{(H,\phi)}(\mathbf{n})$ by $a(\mathbf{n})$ and $A_{(H,\phi)}(\mathbf{x})$ by $A(\mathbf{x})$.

$$a(\mathbf{n}) = \frac{1}{2} \sum_{\mathbf{0}\leq\mathbf{k}\leq\mathbf{n}} \binom{\mathbf{n}}{\mathbf{k}} a(\mathbf{k}) a(\mathbf{n}-\mathbf{k}).$$

The recurrence above is obtained by considering the two subtrees T_1 and T_2, rooted at each of the children of the root of an assembly tree. The tree T_1 is itself an assembly tree of a graph of the form $G_{(H,\phi)}(k_1, k_2, \ldots, k_N)$ where $k_i \leq n_i$ for all i, and T_2 is an assembly tree of a graph of the form $G_{(H,\phi)}(n_1 - k_1, n_2 - k_2, \ldots, n_N - k_N)$. Now

$$
\begin{aligned}
A(\mathbf{x}) &= \sum_{\mathbf{n}\geq\mathbf{0}} a(\mathbf{n}) \frac{\mathbf{x}^{\mathbf{n}}}{\mathbf{n}!} \\
&= \sum_{\mathbf{n}\geq\mathbf{0}} \left\{ \frac{1}{2} \sum_{\mathbf{0}\leq\mathbf{k}\leq\mathbf{n}} \binom{\mathbf{n}}{\mathbf{k}} a(\mathbf{k})a(\mathbf{n}-\mathbf{k}) \right\} \frac{\mathbf{x}^{\mathbf{n}}}{\mathbf{n}!} + \sum_{i=1}^{N} x_i - \sum_{\phi(i)=0} \frac{x_i^2}{2} - \sum_{\{i,j\}\notin E(H)} x_i x_j \\
&= \frac{1}{2} A^2(\mathbf{x}) + \sum_{i=1}^{N} x_i - \sum_{\phi(i)=0} \frac{x_i^2}{2} - \sum_{\{i,j\}\notin E(H)} x_i x_j .
\end{aligned}
$$

The added terms $\sum_{i=1}^{N} x_i - \sum_{\phi(i)=0} \frac{x_i^2}{2} - \sum_{\{i,j\}\notin E(H)} x_i x_j$ correct for the three cases for which the recurrence does not hold. Solving for $A(\mathbf{n})$ by the quadratic formula yields the generating function in the statement of the theorem. □

Some special cases follow as corollaries. For example, if H is just a single vertex v and $\phi(v) = 1$, then $G_{(H,\phi)}(n)$ is the complete graph. Therefore, by Theorem 6 the generating function for the complete graph is $1 - \sqrt{1-2x}$, which when expanded gives the following result.

Corollary 1. *If K_n is the complete graph, then*

$$a(K_n) = \frac{(2n-2)!}{2^{n-1}\,(n-1)!}.$$

Corollary 2. *If $K_{(n_1,n_2,\ldots,n_N)}$ is the complete multipartite graph, then its exponential generating function is*

$$A(\mathbf{x}) = 1 - \sqrt{1 - N + \sum_{i=1}^{N} (1-x_i)^2}.$$

In particular, the generating function for the number of assembly trees of the complete bipartite graph K_{n_1,n_2} is

$$A(x,y) = 1 - \sqrt{(1-x)^2 + (1-y)^2 - 1}.$$

Proof. If H is K_n and ϕ is identically 0 then $G_{(H,\phi)}(n_1, n_2, \ldots, n_N)$ is the complete bipartite graph. □

Expanding for the generating function $A(x, y)$ counting assembly trees of complete bipartite graphs using Maple we arrive at the exponential generating function

$$
\begin{aligned}
A(x,y) = {} & y + x + xy + xy^2 + x^2y + xy^3 + x^3y + (5/2)x^2y^2 + (9/2)x^2y^3 + x^4y \\
& + (9/2)x^3y^2 + xy^4 + 7x^2y^4 + 7x^4y^2 + (25/2)x^3y^3 + 10x^2y^5 + (55/2)x^4y^3 \\
& + (55/2)x^3y^4 + 10x^5y^2 + (27/2)x^2y^6 + (645/8)x^4y^4 + (105/2)x^3y^5 \\
& + (27/2)x^6y^2 + (105/2)x^5y^3 + yx^5 + y^5x + yx^6 + y^6x + yx^7 + y^7x + yx^8 \\
& + (35/2)y^2x^7 + 91y^3x^6 + (1575/8)y^4x^5 + (1575/8)y^5x^4 + 91y^6x^3 \\
& + (35/2)y^7x^2 + y^8x \cdots ,
\end{aligned}
$$

which gives the following table for the number of assembly trees.

1	2	6	24
	10	54	336
		450	3960
			46400

The diagonal elements $1, 10, 450, 23200, \ldots$ are the number of assembly trees for $K_{1,1}, K_{2,2}, K_{3,3}, K_{4,4}, \ldots$, a sequence which does not match anything in the Online Encyclopedia of Integer Sequences [8]. We will return to this sequence in the next section, where we will find the asymptotic growth rate of the sequence, and to find a polynomial recurrence relation satisfied by the sequence.

Consider the set of all (H, ϕ)-graphs on n labeled vertices. In other words, such a graph is obtained by choosing, say n_1 of the n vertices to correspond to vertex 1 of H, n_2 of the n vertices to correspond to vertex 2 of H, ... , n_N of the n vertices to correspond to vertex N of H. Let $b_{(H,\phi)}(n)$ denote the total number of assembly trees of all the possible (H, ϕ)-graphs on n labeled vertices, and $B_{(H,\phi)}(n)$ the corresponding exponential generating function

$$
B_{(H,\phi)}(n) := \sum_{n=0}^{\infty} b_{(H,\phi)}(n) \, \frac{x^n}{n!}.
$$

Corollary 3. *The exponential generating function of the number or assembly trees of all connected (H, ϕ)-graphs of order n such that the number of vertices in H is N, the number of edges in H is M, and the number of i such that $\phi(i) = 0$ is J is*

$$
B_{(H,\phi)}(n) = 1 - \sqrt{1 - 2Nx + \left(2\binom{N}{2} - 2M + J\right) x^2}.
$$

Proof. We have

$$\begin{aligned}
\sum_{i=0}^{\infty} b_{(H,\phi)}(n)\,\frac{x^n}{n!} &= \sum_{n=0}^{\infty}\left[\sum_{n_1+n_1+\cdots+n_N=n}\binom{n}{n_1 n_2\cdots n_N} a_{(H,\phi)}(\mathbf{n})\right]\frac{x^n}{n!} \\
&= \sum_{n=0}^{\infty}\left[\sum_{n_1+n_1+\cdots+n_N=n}\frac{a_{(H,\phi)}(\mathbf{n})}{n_1!n_2!\cdots n_N!}\right] x^n \\
&= \sum_{\mathbf{n}\geq\mathbf{0}} a_{(H,\phi)}(\mathbf{n})\,\frac{x^n}{\mathbf{n}!} \\
&= \sum_{\mathbf{n}\geq\mathbf{0}} a_{(H,\phi)}(\mathbf{n})\,\frac{x^{n_1+\cdots+n_N}}{\mathbf{n}!} \\
&= A_{(H,\phi)}(x,x,\ldots,x) = 1-\sqrt{1-2\sum_{i=1}^{N}x+\sum_{\phi(i)=0}x^2+2\sum_{\{i,j\}\notin E(H)}x^2} \\
&= 1-\sqrt{1-2Nx+\left(2\binom{N}{2}-2M+J\right)x^2},
\end{aligned}$$

where the second-to-last equality follows from Theorem 6. □

4 Examples

In this section, we consider a few interesting examples. In these examples, the graph H is the basis for the construction of an H-graph will be very small (two or three vertices), and $n_1 = n_2$ or $n_1 = n_2 = n_3$ will hold, resulting in graphs $G_{(H,\phi)}$ with two or three classes of vertices in some obvious sense.

4.1 Theoretical Background

4.1.1 Power Series in One Variable

Let $\mathbb{C}[n]$ denote the ring of all polynomials in one variable over the field of complex numbers, and let $\mathbb{C}[[x]]$ denote the ring of all formal power series with complex coefficients. In what follows, we present a few important definitions and theorems on one-variable power series. The interested reader can consult Chapter 6 of [9] for a deeper introduction to the topic, including the proofs of the theorems we include here.

Definition 7. A sequence $f(0), f(1), \cdots$ of complex numbers is called *polynomially recursive*, or *p-recursive* if there exist polynomials $P_0, P_1, \cdots, P_k \in \mathbb{C}[n]$, with $P_k \neq 0$ so that

$$P_k(n+k)f(n+k) + P_{k-1}(n+k-1)f(n+k-1) + \cdots + P_0(n)f(n) = 0 \qquad (1)$$

for all natural numbers n.

Definition 8. We say that the power series $u(x) \in \mathbb{C}[[x]]$ is d-finite if there exists a positive integer d and polynomials $p_0(n), p_1(n), \cdots, p_d(n)$ so that $p_d \neq 0$ and

$$p_d(x)u^{(d)}(x) + p_{d-1}(x)u^{(d-1)}(x) + \cdots + p_1(x)u'(x) + p_0(x)u(x) = 0, \tag{2}$$

Here $u^{(j)} = \frac{d^j u}{dx^j}$.

Theorem 9. *The sequence $f(0), f(1), \cdots$ is p-recursive if and only if its ordinary generating function*

$$u(x) = \sum_{n=0}^{\infty} f(n)x^n \tag{3}$$

is d-finite.

Definition 10. The formal power series $f \in \mathbb{C}[[x]]$ is called *algebraic* if there exist polynomials $P_0(x), P_1(x), \cdots, P_d(x) \in \mathbb{C}[x]$ that are not all equal to zero so that

$$P_0(x) + P_1(x)f(x) + \cdots + P_d(x)f^d(x) = 0. \tag{4}$$

The smallest positive d for which such polynomials exist is called the *degree* of f.

Theorem 11. *If $f \in \mathbb{C}[[x]]$ is algebraic, then it is d-finite.*

We point out that the converse of Theorem 11 is not true. For instance, $f(x) = \sum_{n\geq 1} \frac{x^n}{n} = \ln(1/(1-x))$ is d-finite, but not algebraic, as we will soon see.

One way to prove that a power series is not algebraic is by proving that it is not d-finite. Another way of proving that a power series is not algebraic is by showing that is does not have the "right" growth rate. The following theorem of Jungen is a powerful tool in doing so.

Theorem 12. *[7] Let $f(x) = \sum_{n\geq 0} a_n x^n \in \mathbb{C}[[x]]$ be an* algebraic *power series, and let us assume that $a_n \sim cn^r\alpha^n$, where c and α are* non-zero *complex constants, and r is a* negative real *constant.*

Then $r = s + \frac{1}{2}$, for some negative integer s.

In particular, selecting $c = 1$, $r = -1$ and $\alpha = 1$, we see that $f(x) = \sum_{n\geq 1} \frac{x^n}{n}$ is not algebraic.

4.1.2 Power Series in Several Variables

Now we consider formal power series in several variables. For a deeper introduction to the topic, including the proofs of the theorems we present, see [6]. Let $\mathbb{C}[[x_1, x_2, \cdots, x_k]]$ denote the algebra of all formal power series in variables $x_1, x_2, \cdots, x_k$ over the field of complex numbers.

Definition 13. Let $f(n_1, n_2, \cdots, n_k) : \mathbb{N}^k \to \mathbb{C}$ be a function, and let $F(x_1, x_2, \cdots, x_k) = \sum_{n_1,n_2,\cdots,n_k} f(n_1, n_2, \cdots, n_k)x_1^{n_1}x_2^{n_2}\cdots x_k^{n_k} \in \mathbb{C}[[x_1, x_2, \cdots, x_k]]$.

We say that F is d-finite if all the derivatives

$$\left(\frac{\partial}{\partial x_1}\right)^{d_1}\left(\frac{\partial}{\partial x_2}\right)^{d_2}\cdots\left(\frac{\partial}{\partial x_k}\right)^{d_k}F$$

for $d_i \geq 0$ lay in a finite dimensional vector space over the field of rational functions $\mathbb{C}(x_1, x_2, \cdots, x_k)$.

Theorem 14. *Let $F \in \mathbb{C}[[x_1, x_2, \cdots, x_k]]$. If F is algebraic, then it is d-finite.*

The notion of the *diagonal* of a multivariate power series F is a natural one in that it enables us to focus on the coefficients of F that are often the most interesting for practical purposes.

Definition 15. Let $f(n_1, n_2, \cdots, n_k) : \mathbb{N}^k \to \mathbb{C}$ be a function, and let $F(x_1, x_2, \cdots, x_k) = \sum_{n_1,n_2,\cdots,n_k} f(n_1, n_2, \cdots, n_k)x_1^{n_1}x_2^{n_2}\cdots x_k^{n_k}$. Then the *diagonal* of the multivariate power series $F(x_1, x_2, \cdots, x_k)$ is the *univariate* power series

$$\mathrm{diagF(x)} = \sum_{\mathrm{n}} \mathrm{f(n, n, \cdots, n)x^n}.$$

Example 16. Let $F(s,t) = \frac{1}{1-s-t} = \sum_{m\geq 0}(s+t)^m$. Then for every $n \in \mathbb{N}$, the coefficient of $s^n t^n$ in $F(s,t)$ is equal to $\binom{2n}{n}$. Therefore,

$$\mathrm{diagF(x)} = \sum_{\mathrm{n}} \binom{\mathrm{2n}}{\mathrm{n}}\mathrm{x^n} = \frac{1}{\sqrt{1-\mathrm{4x}}}.$$

Theorem 17. *Let $F \in \mathbb{C}[[x_1, x_2, \cdots, x_k]]$. If F is d-finite, then* diagF(x) *is also d-finite.*

4.2 Three families of graphs

Our first example of computing the diagonal of a power series $A_{(H,\phi)}(\mathbf{z})$ is included because of the precise nature of the answer that we are able to compute.

Example 18. Let H be a graph on vertex set $\{u, v\}$, with $\phi(u) = 0$ and $\phi(v) = 1$, and with one edge, the edge uv. The graphs $G_{(H,\phi)}$ are the graphs consisting of a subgraph G' consisting of n_1 independent vertices and a complete subgraph G'' on n_2 vertices so that G' and G'' are vertex-disjoint, and each vertex of G' is adjacent to each vertex of G''. Then by Theorem 6, we have

$$A_{(H,\phi)}(x, y) = 1 - \sqrt{1 - 2x - 2y + y^2}.$$

In particular, the diagonal of $A_{(H,\phi)}(x, y)$ counts the number of assembly trees of such graphs with $n_1 = n_2$. In order to compute this diagonal, note that

$$A_{(H,\phi)}(x, y) = 1 - \sqrt{1 - 2x - 2y + y^2} = 1 - \left(1 - (2x + 2y - y^2)\right)^{1/2}.$$

By the Binomial theorem, we know that

$$\left((1-(2x+2y-y^2)\right)^{1/2} \;=\; \sum_{m\geq 0}(-1)^m\binom{1/2}{m}\left(2x+2y-y^2\right)^m.$$

When computing the mth power of $(2x+2y-y^2)$, let us consider the summand $(2x)^i(2y)^j(-y^2)^{m-i-j}$. The number of such summands is clearly $\binom{m}{i,j,m-i-j}$. Such a summand will contain x and y raised to the same exponent n if and only if $i=n$ and $n=j+(2m-2n-2j)$, that is, when $3n=2m-j$. In particular, $(2x+2y-y^2)^m$ will contain a constant multiple of x^ny^n if and only if $1.5n\leq m\leq 2n$. Therefore, if we denote the coefficient of x^ny^n in $1-\sqrt{1-2x-2y+y^2}$ by $a_{n,n}$, then routine simplifications lead to the formulas

$$\text{diag}A_{(H,\phi)}(z)=\sum_{n\geq 0}a_{n,n}z^n=\sum_{n\geq 1}\left(\sum_{m=3n/2}^{2n}\binom{1/2}{m}\binom{m}{n}\binom{m-n}{2m-3n}4^{m-n}\right)z^n. \tag{5}$$

and

$$a\left(G_{(H,\phi)}(n,n)\right)=\sum_{m=3n/2}^{2n}\binom{1/2}{m}\binom{m}{n}\binom{m-n}{2m-3n}4^{m-n}. \tag{6}$$

We would like to point out that even if we have given an exact formula for $a_{n,n}$, the question of how fast the $a_{n,n}$ grows is far from being answered. We will discuss that question in Section 4.4. Furthermore, $A_{(H,\phi)}(z)$ is algebraic, so by Theorem 14, it is d-finite. Therefore, Theorem 17 implies that $\text{diag}A_{(H,\phi)}(z)$ is d-finite. So, by Theorem 9, the coefficients of $\text{diag}A_{(H,\phi)}(z)$ must satisfy a polynomial recurrence relation. What is that relation? We will return to this question in Section 4.3.

Example 19. Let H be a graph on vertex set $\{u,v\}$, having one edge, the edge uv, and set $\phi(u)=\phi(v)=0$. Then, as we have seen in Corollary 2, we have

$$A(x,y)=1-\sqrt{(1-x)^2+(1-y)^2-1}$$

for the generating function of the number of assembly trees of a complete bipartite graph.

Determining the coefficients of $\text{diag}A(x,y)$ is much more difficult than it was for the bivariate generating function in Example (18) since directly applying the Binomial theorem would simply lead to formulae that are too complicated to be useful. Nevertheless, some powerful techniques recently developed by Doron Zeilberger and Moa Apagodu will enable us to determine these numbers. We will do this in Section 4.3. The same is true for the case of the complete *tripartite* graph, which is the subject of the next example.

Example 20. Let H be a graph on vertex set $\{u,v,w\}$, having edge set $\{uv,uw,vw\}$, and set $\phi(u)=\phi(v)=\phi(w)=0$. Then, as we have seen in Corollary 2, we have

$$A(x,y,z)=1-\sqrt{(1-x)^2+(1-y)^2+(1-z)^2-2}.$$

4.3 Finding Recurrence Relations

In this section, our goal is to find polynomial recurrence relations for the one-variable generating functions studied in Examples 18, 19 and 20. As we explained in the discussion of Example 18, such recurrence relations exist, since our functions are diagonals of algebraic, and hence, d-finite power series in several variables.

Until recently, the best available technique at this point would have been simply guessing. That is, one would have had to assume that the sought polynomial recurrence relation does not consist of too many terms, and does not involve polynomials of too high degrees, and then have a software package look for a suitable recurrence relation within those limits. One major problem with this approach is that *even if* the software package does return a recurrence relation that is satisfied by all available data points, a theoretical proof that the obtained recurrence relation is satisfied by *all natural numbers* n is still lacking.

The breakthrough is achieved by the following theorem of Zeilberger and Moa. (We present a simplified version of the theorem. The interested reader should consult [2] for the full version.)

Theorem 21. *Let*

$$F(n; x_1, x_2, \cdots, x_d) = \prod_{p=1}^{P} (S_p(x_1, x_2, \cdots, x_d)^{\alpha_p}) \cdot (s(x_1, x_2, \cdots, x_d)t(x_1, x_2, \cdots, x_d))^n$$

where the α_p *are commuting indeterminates, and where the* S_p*,* s *and* t *are elements of* $\mathbb{C}[x_1, x_2, \cdots, x_d]$.

Then there exists a non-negative integer L*, there exist* $L+1$ *polynomials in* n*,* $e_0(n), e_1(n), \cdots, e_L(n)$*, not all zero, and there exist* d *rational functions* $R_i(n; x_1, x_2, \cdots x_d)$ *(with* $i = 1, 2, \cdots, d$*) such that the functions*

$$G_i(n; x_1, x_2, \cdots x_d) := R_i(n; x_1, x_2, \cdots x_d)F(n; x_1, x_2, \cdots, x_d)$$

satisfy the equations

$$\sum_{i=0}^{L} e_i(n)F(n+i; x_1, x_2, \cdots, x_d) = \sum_{i=1}^{d} D_{x_i} G_i(n; x_1, x_2, \cdots x_d). \qquad (7)$$

Furthermore, there exists a constant $N = N(deg(s), deg(t), \sum_{p=1}^{P} deg(S_p))$ *such that* $L \leq N$ *and* $deg(e_i) \leq N$*, and therefore, the polynomials* $e_i(n)$ *can be explicitly computed.*

Note that Theorem 21 eliminates the need for "guessing" the polynomial recurrence satisfied by the functions $F(n+i; x_1, x_2, \cdots, x_d)$. Indeed, because of the existence of the upper bound N for the number and maximum degree of the polynomials $e_i(n)$, finding the polynomials $e_i(n)$ in (7) is simply the question of solving a (possibly huge) system of linear equations. Furthermore, by Theorem 21 we know that a polynomial recurrence relation (7) exists, so once we have enough equations in our system to obtain a *unique* solution for the vector $e(n) = (e_0(n), e_1(n), \cdots, e_L(n))$, we can be sure that the polynomial recurrence relation defined by $e(n)$ is indeed a recurrence relation that is satisfied by the $F(n+i; x_1, x_2, \cdots, x_d)$ for *all* n. So Theorem 21 does take care of both problems we had with simply guessing a polynomial recurrence relation.

Now apply Theorem 21 to Example 18 by letting $d = 2$, $x_1 = x$, and $x_2 = y$. Set $S_1(x, y) = 1 - 2x - 2y + y^2$, with $\alpha_1 = 1/2$, and, crucially, $S_2(x, y) = xy$ with $\alpha_2 = -1$. Finally, set $s(x, y) = 1$ and $t(x, y) = xy$. This leads to

$$F(n; x, y) = \frac{\sqrt{1 - 2x - 2y + y^2}}{x^{n+1} y^{n+1}}. \tag{8}$$

Consider the right-hand side as a power series in two variables and integrate both sides of (8) on a two-dimensional polydisk whose interior contains 0. By the two-variable residue theorem, we see that the only summand in the right-hand side whose integral does not vanish is $a_{n,n} \frac{1}{x} \cdot \frac{1}{y}$. In fact, by the residue theorem, we get

$$\int F(n, x, y) = -4\pi^2 a_{n,n}.$$

Therefore, the polynomial recurrence relation (7) is equivalent to a polynomial recurrence relation for the numbers $a_{n,n}$. In order to obtain this recurrence relation, we need to solve a large system of linear equations. Fortunately, Doron Zeilberger's software package, SMAZ [3] can do that for us. The result is the following theorem.

Theorem 22. *Let $a_n = a_{n,n}$ be the coefficient of z^n in (5). Recall that a_n counts assembly trees of graphs studied in Example 18. Then we have $a_0 = 0$, $a_1 = 1$, and*

$$a_{n+1} = \frac{3}{2} \cdot \frac{(3n-1)(3n+1)}{(n+1)^2} a_n. \tag{9}$$

(Note that the above discussion computes a recurrence relation for the coefficients of $\mathrm{diag}\sqrt{1 - 2\mathrm{x} - 2\mathrm{y} + \mathrm{y}^2}$ and not $\mathrm{diag}(1 - \sqrt{1 - 2\mathrm{x} - 2\mathrm{y} + \mathrm{y}^2})$, but it is obvious that starting with the coefficient of z, the coefficients of these two power series will satisfy the same recurrence relation.)

Similarly, we can apply Theorem 21 to obtain a polynomial recurrence relation for the coefficients b_n of z^n in $\mathrm{diag}A(z)$, where $A(x, y)$ is the bivariate generating function of the number of assembly trees of complete bipartite graphs as computed in Example 19. We assign the same values to the various parameters as we did immediately preceding (8), except that we set $S_1(x, y) = x^2 + y^2 - 2x - 2y + 1$. The result is the following.

Let $[z^n]g(z)$ denote the coefficient of z^n in the power series $g(z)$.

Theorem 23. *Let $b_n = [z^n]\mathit{diag}\left(1 - \sqrt{x^2 + y^2 - 2x - 2y + 1}\right)$. Note that b_n is the number of assembly trees of $K_{n,n}$, the graph studied in Example 19. Then we have $b_0 = 0$, $b_1 = 1$, $b_2 = 5/2$, and*

$$b_{n+2} = \frac{2(6n^2 + 12n + 5)}{(n+2)^2} b_{n+1} - \frac{n(2n-1)(2n+3)}{(n+2)^2(n+1)} b_n \tag{10}$$

if $n \geq 1$.

Finally, let $c_n = [t^n]\mathrm{diag}A(t)$, where $A(x,y,z)$ is the trivariate generating function of the number of assembly trees of complete tripartite graphs as computed in Example 20. We can then use Theorem 21 with $d=3$, $x_1 = x$, $x_2 = y$, $x_3 = z$. Set $S_1(x,y,z) = A(x,y,z)$ as defined in Example 20 with $\alpha_1 = 1/2$, and $S_2(x,y,z) = xyz$ with $\alpha_2 = -1$. Finally, set $s(x,y,z) = 1$, and $t(x,y,z) = xyz$. We then get the following result.

Theorem 24. *Let $c_n = [t^n]diagA(t)$, where*

$$A(x,y,z) = 1 - \sqrt{(1-x)^2 + (1-y)^2 + (1-z)^2 - 2}.$$

Then we have $c_0 = 0$, $c_1 = 3$, $c_2 = 84$, $c_3 = 4935$, and

$$c_{n+3} = r_2(n)c_{n+2} + r_1(n)c_{n+1} + r_0(n)c_n, \tag{11}$$

if $n \geq 1$. Here the r_i are explicitly known rational functions of n, with numerators and denominators of degree 11 for r_0, degree ten for r_1, and degree nine for r_2.

4.4 Finding Growth Rates

As far as determining the growth rate of the sequence $a_1, a_2, \cdots$, recurrence relation (9) is much more useful than the explicit formula that we found for a_n in (5). Indeed, the *exponential* growth rate of a_n is easy to read off from (9). It is routine to prove that

$$\lim_{n\to\infty} \sqrt[n]{a_n} = 13.5.$$

Determining the growth rate of a_n at a higher level of precision is much more difficult. The theoretical foundation of this computation is the paper [10] by Doron Zeilberger and Jet Wimp. In that survey paper, the authors discuss results of Birkhoff and Trzinski in a more general setup, but in the examples we study, their method simplifies to the following.

Let t_n be a sequence for which a polynomial recurrence relation is known, and of which we want to compute the asymptotics. Try to obtain t_n in the form

$$t_n = E_n K_n,$$

where

$$E_n = e^{\mu_0 n \ln n + \mu_1 n \ln n} n^\theta,$$

and

$$K_n = \exp\left(\alpha_1 n^\beta + \alpha_2 n^{\beta - \frac{1}{\rho}} + \alpha_3 n^{\beta - \frac{2}{\rho} + \cdots}\right).$$

Here $\alpha_1 \neq 0$, $\beta = j/\rho$, and $0 < j < \rho$. This decomposition leads to the formula

$$\frac{t_{n+k}}{t_n} = n^{\mu_0 k}\lambda^k\left(1 + \frac{k\theta + k^2\mu_0/2}{n} + \cdots\right)\cdot\exp\left(\alpha_1 n^\beta k n^{\beta-1} + \alpha_2\left(\beta - \frac{1}{\rho}k\right)n^{\beta-1-\frac{1}{\rho}} + \cdots\right),$$

where $\lambda = e^{\mu_0 + \mu_1}$. Then, using the polynomial recurrence relation for t_n, determine the parameters in E_n and K_n, obtaining this way an *exact* formula (in the form of an infinite sum) for t_n. This computation can certainly be long and teadious, but the software package AsyRec [11] can carry it out. For the sequence a_n, we obtain the following result.

Theorem 25. *Let a_n be defined as in Theorem 22. Note that a_n is the number of assembly trees of the graph discussed in Example 18. Then we have*

$$a_n = C \cdot \frac{13.5^n}{n^2} \cdot \left(1 + \frac{1}{9n} + \frac{5}{81n^2} + \cdots\right),$$

where C is an absolute constant. In particular, $a_n \sim \frac{13.5^n}{n^2}$, and therefore, by Theorem 12, $\mathrm{diagA}_{(\mathrm{H},\phi)}(\mathrm{z})$ *is* not *algebraic.*

We do not know of a theoretical method to determine the exact value of C, but an estimate can be obtained by using the command AsyC in the package AsyRec [11]. In this example, the program returns $C = 0.0612587....$

Applying the same method for the polynomial recurrence relation proved for the numbers b_n in Theorem 23, we get the following asymptotic expressions.

Theorem 26. *Let b_n be defined as in Theorem 23, that is, let b_n be the number of assembly trees of the complete bipartite graph $K_{n,n}$, which was studied in Example 19. Then we have*

$$b_n = C \cdot \frac{(6+4\sqrt{2})^n}{n^2} \left(1 + \frac{35}{8n} - \frac{5}{32}\frac{\sqrt{2}}{n} + \cdots\right).$$

In particular, $b_n \sim \frac{(6+4\sqrt{2})^n}{n^2}$, and therefore, by Theorem 12, $\mathrm{diagA}_{\mathrm{x,y}}(\mathrm{z})$ *is* not *algebraic.*

Again, C is an absolute constant whose exact value we were not able to determine. The estimate we get from AsyC is $C = 0.0717565....$

5 Questions

There are reasonable alternatives to the edge gluing rule for defining an assembly tree. Two possibilities are the following. For each of these two rules, the problem is again to enumeratate the number of assembly trees for interesting graphs.

Definition 27. An assembly tree for a connected graph G using the *connected gluing rule* is an assembly tree for G (satisfying properties (1-4) in Section 1), and also satisfying the additional property:

5. For each node, the graph induced by the vertices in the label is connected.

This rule is less restrictive than the edge gluling rule. In particular, an assembly tree of a graph G is not necessarily a binary tree. At the extreme is the assembly tree of depth 1 for which every vertex of G is a child of the root.

Definition 28. In this definition, we denote each face of a plane graph by the set of vertices on that face. An assembly tree for a connected plane graph G using the *face gluing rule* is an assembly tree for G satisfying the additional property.

5. For each internal node U there is a face F such that $C \cap F \neq \emptyset$ for each $C \in c(U)$ and $F \subseteq \bigcup c(U)$.

Figure 5 shows a plane graph and one assembly tree using the face gluing rule.

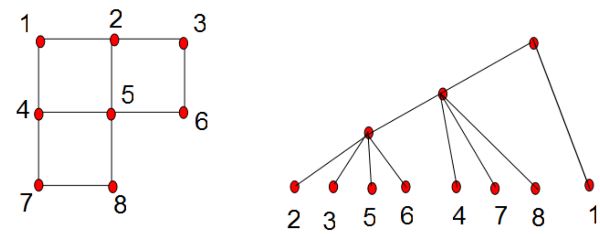

Figure 4: A plane graph and an assembly tree using the face gluing rule.

All the graphs in Sections 2 and 3 for which we were able to compute the number of assembly trees share a common property: the number of connected induced subgraphs, up to unlabeled isomorphism type, is small. For example, the number of connected induced subgraphs of the path P_n is n. For the complete graph K_n, it is n, and for the complete bipartite graph $K_{m,n}$ it is mn. With (H, ϕ) fixed, the number of connected induced subgraphs of any (H, ϕ)-graph is again polynomial in the parameters $n_1, n_2, \ldots, n_N$. For results on graphs with few isomorphism types of induced subgraphs see [1] and [5]. So the question arises: is it possible to enumerate the number of assembly trees (by the edge gluing rule) for a family of graphs for which the number of isomorhism types of induced subgraphs is not small. In particular, consider the family of caterpillar graphs D_n on $2n$ vertices as shown in Figure 5. The number of isomorphism types of linduced subgraphs is clearly large, exponential in n.

Question. Is there a reasonable enumeration of the number of assembly trees for the family D_n?

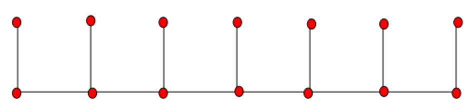

Figure 5: Caterpillar graph D_7.

Acknowledgment

We are indebted to Doron Zeilberger, who has made us aware of his results in [2], and has patiently explained his method to us. We are also grateful to Richard Stanley, Leonard Lipshitz, Robin Pemantle and Vincent Vatter for valuable suggestions.

References

[1] N. Alon and B. Bollobás, Graphs with a small number of distinct induced subgraphs, *Disc. Math*, **75** (1989) 23-30.

[2] M. Apagodu, D. Zeilberger, Multi-variable Zeilberger and Almkvist-Zeilberger algorithms and the sharpening of Wilf-Zeilberger theory, *Adv. Appl. Math.* **37** (2006), 139–152.

[3] M. Apagodu, D. Zeilberger, Multi-Variable Zeilberger and Almkvist-Zeilberger Algorithms and the Sharpening of Wilf-Zeilberger Theory, software package, available at `http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/multiZ.html`.

[4] M. Bóna, M. Sitharam, A. Vince, Tree orbits under permutation group action: algorithm, enumeration and application to viral assembly, *Bull. Math. Biology*, **73** (2011) 726-753.

[5] P. Erdős and A. Hajnal, On the number of distinct induced subgraphs of a graph, *Disc. Math*, **75** (1989) 145-154.

[6] L. Lipshitz, *D*-finite power series, *J. Algebra* **122** (1989) 353–373.

[7] R. Jungen, Sur les séries de Taylor n'ayant que des singularités algébro-logarithmiques sure leur cercle de convergence, *Comment. Math. Helv.* **3** (1931), 266–306.

[8] The Online Encyclopedia of Integer Sequences, *internet resource*, `https://oeis.org/`.

[9] R. Stanley, Enumerative Combinatorics, Volume 2, *Cambridge University Press*, Cambridge UK, 1999.

[10] J. Wimp, D. Zeilberger, Resurrecting the Asymptotics of Linear Recurrences. *J. Math. Anal. Appl.* **111** (1985), 162–176.

[11] . D. Zeilberger, AsyRec, a Maple package that automatically computes the asymptotics for solutions of linear recurrence eqations with polynomial coefficients. Available at `http://www.math.rutgers.edu/~zeilberg/tokhniot/AsyRec`.