

COUNTING MARKOV EQUIVALENCE CLASSES FOR DAG MODELS ON TREES

ADITYANARAYANAN RADHAKRISHNAN, LIAM SOLUS, AND CAROLINE UHLER

ABSTRACT. DAG models are statistical models satisfying a collection of conditional independence relations encoded by the nonedges of a directed acyclic graph (DAG) $\mathcal{G}$. Such models are used to model complex cause-effect systems across a variety of research fields. From observational data alone, a DAG model $\mathcal{G}$ is only recoverable up to *Markov equivalence*. Combinatorially, two DAGs are Markov equivalent if and only if they have the same underlying undirected graph (i.e. skeleton) and the same set of the induced subDAGs $i \rightarrow j \leftarrow k$, known as immoralities. Hence it is of interest to study the number and size of Markov equivalence classes (MECs). In a recent paper, the authors introduced a pair of generating functions that enumerate the number of MECs on a fixed skeleton by number of immoralities and by class size, and they studied the complexity of computing these functions. In this paper, we lay the foundation for studying these generating functions by analyzing their structure for trees and other closely related graphs. We describe these polynomials for some important families of graphs including paths, stars, cycles, spider graphs, caterpillars, and complete binary trees. In doing so, we recover important connections to independence polynomials, and extend some classical identities that hold for Fibonacci numbers. We also provide tight lower and upper bounds for the number and size of MECs on any tree. Finally, we use computational methods to show that the number and distribution of high degree nodes in a triangle-free graph dictates the number and size of MECs.

1. INTRODUCTION

A graphical model based on a directed acyclic graph (DAG), known as a *DAG model* or *Bayesian network*, is a type of statistical model used to model complex cause-and-effect systems. DAG models are popular in numerous areas of research including computational biology, epidemiology, environmental management, and sociology [1, 17, 35, 40, 43]. Given a DAG $\mathcal{G} := ([p], A)$ with nodes $[p] = \{1, \dots, p\}$ and arrows $i \rightarrow j \in A$, the DAG model associates to each node $i \in [p]$ of $\mathcal{G}$ a random variable X_i . The collection of non-arrows of $\mathcal{G}$ encode those conditional independence (CI) relations typical of cause-effect relationships:

$$X_i \perp\!\!\!\perp X_{\text{nd}(i) \setminus \text{pa}(i)} \mid X_{\text{pa}(i)},$$

where $\text{nd}(i)$ and $\text{pa}(i)$ respectively denote the *nondesendents* and *parents* of the node i in $\mathcal{G}$. A probability distribution $\mathbb{P}$ is said to satisfy the *Markov assumption* with respect to $\mathcal{G}$ if it entails these CI relations, and the DAG model associated to $\mathcal{G}$ is the complete set of all such joint probability distributions. The global consequences of the Markov assumption in terms of CI relations can be captured via the combinatorics of the DAG $\mathcal{G}$ with a notion of directed separation called *d-separation* [14, Chapter 3]. Unfortunately, multiple DAGs can encode the same

set of CI relations. Such DAGs are said to be *Markov equivalent*, and the complete collection of DAGs encoding the same set of CI relations as $\mathcal{G}$ is called the *Markov equivalence class* (MEC) of $\mathcal{G}$. Verma and Pearl show in [49] that a MEC is combinatorially determined by the underlying undirected graph G (or *skeleton*) of $\mathcal{G}$ and the placement of immoralities, i.e. induced subgraphs of the form $i \rightarrow j \leftarrow k$.

From observational data, the underlying DAG $\mathcal{G}$ of a DAG model can only be determined up to Markov equivalence. It is therefore of interest to gain a combinatorial understanding of MECs, in particular their number and sizes. The literature on the MEC enumeration problem can be summarized via the following three perspectives: (1) count the number of MECs on all DAGs on p nodes [20], (2) count the number of MECs of a given size [19, 45, 50], or (3) determine the size of a specific MEC [22, 23]. In [20], the authors approach perspective (1) computationally and compute the number of MECs for all DAGs on $p \leq 10$ nodes. In [19, 45, 50], the authors provide partial results for perspective (2) using inclusion-exclusion formulae that work nicely for small MECs sizes. Then in [22, 23], the authors explore efficient techniques for computing the size of a fixed MEC via algorithms that manipulate *v-rooted* and *core* subgraphs of chordal graphs. Recently, [39] addresses this question from a new perspective by introducing a pair of generating functions that enumerate the number of MECs on a *fixed skeleton* $G = (V, E)$ by number of immoralities in each class and by class size. Their results reveal connections to graphical enumeration problems that are well-studied from the perspective of combinatorial optimization. A main goal of this paper is make explicit these connections and use them to study the generating functions of [39] for sparse graphs.

Throughout, we use curly letters for DAGs, such as $\mathcal{G}$, and script letters for the corresponding undirected graph (i.e. skeleton), such as G . In addition, we use A to denote a collection of arrows and E to denote a collection of undirected edges. The first generating function is the graph polynomial

$$M(G; x) := \sum_{k \geq 0} m_k(G) x^k,$$

where $m_k(G)$ denotes the number of MECs with skeleton G that contain precisely k immoralities. The degree of $M(G; x)$, denoted $m(G)$, is called the *immorality number* of G , and it counts the maximum number of immoralities possible in an MEC with skeleton G . The second generating function is the arithmetic function

$$S(G; x) := \sum_{k \geq 0} \frac{s_k(G)}{k^x},$$

where $s_k(G)$ denotes the number of MECs with skeleton G that have size k . We let $M(G) := M(G; 1) = S(G; 0)$ denote the total number of MECs with skeleton G . In [39], the authors showed that computing a DAG with $m(G)$ immoralities is an NP-hard problem, and that $S(G; x)$ is a complete graph isomorphism invariant for all connected graphs on $p \leq 10$ nodes. Otherwise, very little is known about the structure of these generating functions.

In this paper, we lay the foundation for the study of the graph polynomial $M(G; x)$ by providing a detailed analysis of its properties for trees (and their closely related graphs). Within this context, we draw explicit connections between properties of $M(G; x)$ and the *independence polynomial* of G ; i.e. the graph polynomial $I(G; x) := \sum_{k \geq 0} \alpha_k(G) x^k$, where $\alpha_k(G)$ denotes the number of pairwise disjoint k -subsets of vertices (*independent sets*) of G .

The remainder of this paper is structured as follows: In Section 2 we compute $M(G; x)$ and $S(G; x)$ for some fundamental examples, including paths, cycles, and stars. We find that $M(G; x)$ coincides with an independence polynomial for paths and cycles, therein providing connections to Fibonacci numbers and Fibonacci-like sequences. Paths and stars give tight bounds on the number of independence sets in a tree [37]. We show in Section 3 that they also provide tight upper and lower bounds for the number and sizes of MECs on a tree. In Section 4 we then use $M(G; x)$ for stars and paths to compute $M(G; x)$ and $M(G)$ for families of trees that are significant in both mathematical and statistical settings. The graphs analyzed include spider graphs, caterpillar graphs, and complete binary trees. In the case of spider graphs, the resulting formulae yield generalizations of classic identities known for Fibonacci numbers, and reveal a multivariate extension of $M(G; x)$ exhibiting nice combinatorial properties that can be recursively computed for any tree. In Section 5, we use computational methods to examine properties of $M(G; x)$ and $M(G)$ for the more general family of triangle-free graphs. The results of [39] and those of Sections 2, 3, and 4 exhibit an underlying relationship between the number and size of MECs and the number of cycles and high degree nodes in the graph. Using a program first described in [39], we study this connection by examining data collected on MECs for all connected graphs on $p \leq 10$ nodes. We compare class size and the number of MECs per skeleton to skeletal features including average degree, maximum degree, clustering coefficient, and the ratio of number of immoralities in the MEC to the number of induced 3-paths in the skeleton. Unlike $S(G; x)$, the polynomial $M(G; x)$ is not a complete graph isomorphism invariant over all connected graphs on $p \leq 10$ nodes. However, using this program, we observe that it is such an invariant when restricted to triangle-free graphs.

2. SOME FIRST EXAMPLES

In this section, we compute the generating functions $M(G; x)$ and $S(G; x)$ for paths, cycles, stars, and bistars. We show that $M(G; x)$ are independence polynomials for all paths and cycles. Similarly, we show that for the star graphs $M(G; x)$ has nonzero coefficients given by the binomial coefficients, which are precisely the coefficients of its corresponding independence polynomial. These examples are fundamental to the theory developed in Sections 3 and 4, in which we bound the number and size of MECs on trees and compute $M(G; x)$ for more general families of graphs using paths and stars.

Recall that the p -path is the (undirected) graph $I_p := ([p], E)$ for which $E := \{\{i, i+1\} : i \in [n-1]\}$, and the p -cycle is the (undirected) graph $C_p := ([p], E)$ for which $E := \{\{i, i+1\} : i \in [n-1]\} \cup \{\{1, n\}\}$. We also define the graph $G_p(q_1, q_2, \dots, q_p)$ to be the undirected graph given by attaching q_i leaves to node i of the p -path I_p . The p -star is the graph $G_1(p)$ and the p, q -bistar is the graph $G_2(p, q)$. The *center* node of $G_1(p)$ is its unique node of degree p .

2.1. Paths and cycles. We introduce two well-studied combinatorial sequences, and their associated polynomial filtrations that will play a fundamental role in the formulae computed in this section as well as in Sections 3 and 4. Recall that the p^{th} Fibonacci number F_p is defined by the recursion

$$F_0 := 1 \quad F_1 := 1, \quad \text{and} \quad F_p := F_{p-1} + F_{p-2} \quad \text{for } p \geq 2.$$

The p^{th} *Fibonacci polynomial* is defined by

$$F_p(x) := \sum_{k=0}^{\lfloor \frac{p}{2} \rfloor} \binom{p-k}{k} x^k,$$

and it has the properties that $F_p(1) = F_p$ for all $p \geq 1$ and $F_p(x) = F_{p-1}(x) + xF_{p-2}(x)$ for all $p \geq 2$. Analogously, the p^{th} Lucas number L_p is given by the Fibonacci-like recursion

$$L_0 := 2 \quad L_1 := 1, \quad \text{and} \quad L_p := L_{p-1} + L_{p-2} \quad \text{for } p \geq 2.$$

The p^{th} *Lucas polynomial* is given by

$$L_0(x) := 2 \quad L_1(x) := 1, \quad \text{and} \quad L_p(x) := L_{p-1}(x) + xL_{p-2}(x) \quad \text{for } p \geq 2.$$

It is a well-known that the independence polynomial of the p -path is equal to the $(p+1)^{\text{st}}$ Fibonacci polynomial and the independence polynomial of the p -cycle is given by the p^{th} Lucas polynomial; i.e.

$$I(I_p; x) = F_p(x) \quad \text{and} \quad I(C_p; x) = L_p(x).$$

With these facts in hand we prove the following theorem.

Theorem 2.1. *For the path I_p and the cycle C_p on p nodes we have that*

$$M(I_p; x) = F_{p-1}(x) \quad \text{and} \quad M(C_p; x) = L_p(x) - 1.$$

In particular, the number of MECs on I_p and C_p , respectively, is

$$M(I_p) = F_{p-1} \quad \text{and} \quad M(C_p) = L_p - 1,$$

and the maximum number of immoralities is

$$m(I_{p+2}) = m(C_p) = \left\lfloor \frac{p}{2} \right\rfloor.$$

Proof. The result follows from a simple combinatorial bijection. Since paths and cycles are the graphs with the property that the degree of any vertex is at most two, then the possible locations of immoralities are exactly the degree two nodes. That is, the unique head node j in an immorality $i \rightarrow j \leftarrow k$ must be a degree two node. In the path I_p , this corresponds to all $p-2$ non-leaf vertices, and for the cycle C_p this is all the vertices of the graph. Notice then that no two adjacent degree two nodes can simultaneously be the unique head node of an immorality, since this would require one arrow to be bidirected. Thus, a viable placement of immoralities corresponds to a choice of any subset of degree two nodes that are mutually non-adjacent, i.e. that form an independent set.

Conversely, given any independent set in I_p , a DAG can be constructed by placing the head node of an immorality at each element of the set and directing all other arrows in one direction. Similarly, this works for any nonempty independent set in C_p . (Notice that any MEC on the cycle must have at least one immorality since all DAGs have at least one sink node.) The resulting formulas are then

$$M(I_p; x) = I(I_{p-2}; x) = F_{p-1}(x) \quad \text{and} \quad M(C_p; x) = I(C_p; x) - 1 = L_p(x) - 1,$$

which completes the proof. $\square$

We now compute the generating functions $S(I_p; x)$ and $S(C_p; x)$. The desired formulae follow naturally from the description of the placement of immoralities given in Theorem 2.1.

Theorem 2.2. *The number $s_\ell(I_p)$ of MECs of size ℓ with skeleton I_p is the number of compositions $c_1 + \dots + c_{k+1} = p - k$ of $p - k$ into $k + 1$ parts that satisfy*

$$\ell = \prod_{i=1}^{k+1} c_i$$

as k varies from $0, 1, \dots, \lfloor \frac{p}{2} \rfloor$.

Proof. Let $\mathcal{G}$ be a DAG with skeleton I_p . We denote the Markov equivalence class of $\mathcal{G}$ by $[\mathcal{G}]$. By the proof of Theorem 2.1, we know that the immorality placements in $[\mathcal{G}]$ correspond to the nodes in an independent k -subset $\mathcal{I} \subset [p]$ on the subpath I_{p-2} of I_p induced by the non-leaf nodes of I_p . The induced graph of the complement of $\mathcal{I}$ is a forest of $k + 1$ paths. Since each member of $[\mathcal{G}]$ is a DAG with skeleton I_p that has no immoralities on these $k + 1$ paths, then each path contains a unique sink. Each independent k -subset yields a distinct forest of $k + 1$ paths on $[p] \setminus \mathcal{I}$, which corresponds to a unique partition of $p - k$ into $k + 1$ parts. The formula for $s_\ell(I_p)$ is then given by considering all such possible placements of sinks on each path in the forests over all independent sets. $\square$

A similar argument using integer partitions allows us to compute the number of MECs of size ℓ on the p -cycle.

Theorem 2.3. *The number of MECs of size ℓ in the p -cycle is*

$$s_\ell(C_p) = \sum_{k=1}^{\lfloor \frac{p}{2} \rfloor} \sum_{\substack{\mathbf{m} \in \mathbb{P}[p-2k+1, k, p-k], \\ \ell = \prod_{i=1}^k i^{m_i}}} \frac{p}{k} \binom{k}{m_1, \dots, m_{p-2k+1}},$$

where $\mathbb{P}[j, k, n]$ denotes the partitions of n with k parts with largest part at most j .

Proof. Since C_p is a graph in which every node is degree 2, then each MEC of C_p containing k immoralities corresponds to an independent k -subset of $[p]$, and the subgraph of C_p given by deleting this k -subset consists of k disjoint paths. The size of this MEC is then the product of the lengths of these paths. So we need only count the number of such subgraphs for which this product equals ℓ .

To count these objects, consider that each subgraph of C_p given by deleting an independent k -subset of C_p forms a partition of the $p - k$ remaining vertices into k parts with maximum possible part size being $p - 2k + 1$. Such a partition is represented by

$$\langle 1^{m_1}, 2^{m_2}, \dots, (p - 2k + 1)^{m_{p-2k+1}} \rangle \in \mathbb{P}[p - 2k + 1, k, p - k],$$

where $m_1, \dots, m_{p-2k+1} \geq 0$ and $\sum_i m_i = k$. Each such partition corresponds to an unlabeled forest consisting of m_i i -paths, and the number of subgraphs of C_p isomorphic to this forest is

$$\frac{p}{k} \binom{k}{m_1, \dots, m_{p-2k+1}}.$$

The claim follows since the size of each corresponding MEC is $\prod_{i=1}^k i^{m_i}$. $\square$

Remark 2.1. It is a well-known result that the coefficient of x^k in the $(p - 1)^{st}$ Fibonacci polynomial is the binomial coefficient $\binom{p-k-1}{k}$, and that this is also the number of compositions of $p - k$ into $k + 1$ parts. The former result says that the

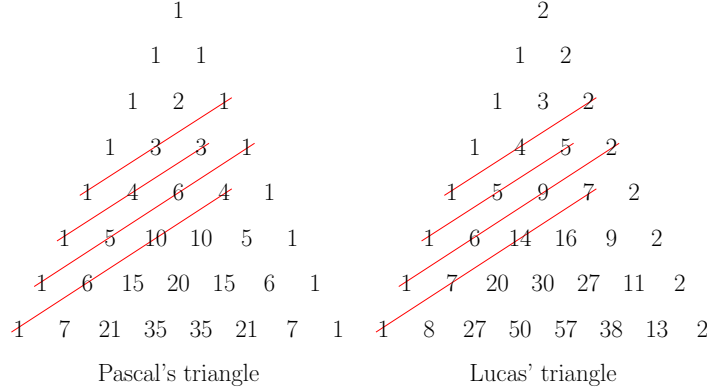


FIGURE 1. The p^{th} diagonal of each triangle is the coefficient vector of $F_p(x)$ and $L_p(x)$, respectively.

$(p-1)^{st}$ Fibonacci polynomial has coefficients given by the $(p-1)^{st}$ diagonal of Pascal's triangle, and so the latter result gives a compositional interpretation of the corresponding entry in Pascal's Triangle; see Figure 1 (left). In this section, we saw that this compositional interpretation of $\binom{p-k-1}{k}$ results in the proof of Theorem 2.2.

Analogously, the p^{th} diagonal of a second triangle, called *Lucas' triangle* in [7], corresponds to the coefficients of the p^{th} Lucas polynomial. This triangle is depicted on the right in Figure 1. Thus, the proof of Theorem 2.3 results in a combinatorial interpretation of the entries of this triangle via partitions. In particular, the entry of the Lucas triangle corresponding to the k^{th} coefficient of $L_p(x)$ is

$$[x^k].L_p(x) = \sum_{\mathbf{m} \in \mathbb{P}[p-2k+1, k, p-k]} \frac{p}{k} \binom{k}{m_1, \dots, m_{p-2k+1}}.$$

Moreover, the binomial recursion on the triangle implies that these coefficients satisfy the identity

$$[x^k].L_p(x) = [x^{k-1}].L_{p-2} + [x^k].L_{p-1}.$$

To the best of the authors' knowledge, such a partition identity is new to the combinatorial literature.

2.2. Stars and bistars. We now study the star and bistar graphs, $G_1(p)$ and $G_2(p, q)$. An example of a star and a bistar is given in Figure 2. The number of

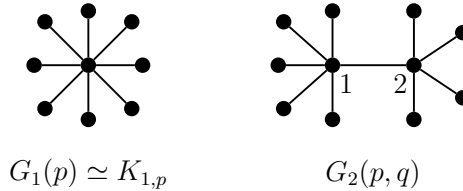


FIGURE 2. On the left is a star and on the right is a bistar.

MECs on stars and their sizes will play an important role in Sections 3 and 4.

Theorem 2.4. *The MECs on the p -star $G_1(p)$ have the polynomial generating function*

$$M(G_1(p); x) = 1 + \sum_{k \geq 2} \binom{p}{k} x^{\binom{k}{2}}.$$

In particular,

$$M(G) = 2^p - p.$$

Moreover, the corresponding class sizes are

$$s_1(G_1(p)) = 2^p - p + 1 \quad \text{and} \quad s_{p+1}(G_1(p)) = 1.$$

Proof. Any immorality $i \rightarrow j \leftarrow k$ in a DAG on $G_1(P)$ must have the unique head node j being the center node of $G_1(P)$, and the tail nodes i and k must be leaves of $G_1(p)$. It follows that each MEC on $G_1(p)$ having at least one immorality is given by selecting any k -subset of the p leaves for $k \geq 2$ to be directed towards the center node and then directing all other edges outwards. Each such k -subset yields a unique MEC of size one containing $\binom{k}{2}$ immoralities. The final MEC is the class containing no immoralities. This class consists of all DAGs on $G_1(p)$ with a unique source node, and there are $p + 1$ such DAGs. $\square$

The formulas in Theorem 2.4 allow us to obtain similar formulas for bistars. For convenience, we let

$$P_m := \sum_{k=1}^m \binom{m}{k} x^{\binom{k+1}{2}}.$$

It will also be helpful to label edges that have specified roles in certain MECs. The green edges (also labeled with $\square$) indicate that these edges cannot be involved in any immorality. The red arrows (also labeled with $*$) indicate a fixed immorality in the partially directed graph, and the blue arrows (also labeled with $\circ$) represent fixed arrows that are not in immoralities.

Theorem 2.5. *The MECs on the bistar $G_2(p, q)$ have the polynomial generating function*

$$M(G_2(p, q); x) = M(G_1(p); x)P_q + M(G_1(q); x)P_p + M(G_1(p); x) + M(G_1(q); x) - 1.$$

In particular,

$$M(G_1(p, q)) = 2^{p+q+1} - p2^q - q2^p - 1.$$

Moreover, the corresponding class sizes are

$$s_1(G_2(p, q)) = 2^{p+q+1} - p2^q - q2^p - 2^p - 2^q,$$

$$s_{p+1}(G_2(p, q)) = 2^q - 1, \quad s_{q+1}(G_2(p, q)) = 2^p - 1, \quad \text{and} \quad s_{p+q+2}(G_2(p, q)) = 1.$$

Proof. To count the MECs on the bistar $G_2(p, q)$ we consider three separate cases defined in terms of the edge $\{1, 2\}$. These three cases are:

- (1) The edge $\{1, 2\}$ is in an immorality with at least one of the p leaves attached to node 1.
- (2) The edge $\{1, 2\}$ is in an immorality with at least one of the q leaves attached to node 2.
- (3) The edge $\{1, 2\}$ is not in an immorality.

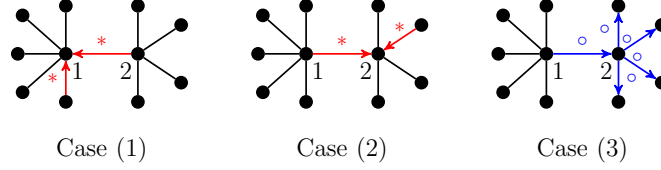


FIGURE 3. The three cases of the proof of Theorem 2.5.

The three cases are depicted in Figure 3. In the first case, at least one of the p leaves attached to node 1 must be in an immorality with the edge $\{1, 2\}$, and the q leaves attached to node 2 can display any pattern of immoralities of the star $G_1(q)$. This yields $M(G_1(q); x)P_p$ MECs as counted by their number of immoralities. Similarly, case two yields $M(G_1(p); x)P_q$. In the third case, in order for the edge $\{1, 2\}$ to not appear in any immorality, we need that all edges at the head of $\{1, 2\}$ point towards the leaves. This yields $M(G_1(p); x) + M(G_1(q); x) - 1$ MECs as counted by their number of immoralities. Thus,

$$M(G_2(p, q); x) = M(G_1(p); x)P_q + M(G_1(q); x)P_p + M(G_1(p); x) + M(G_1(q); x) - 1,$$

and evaluating this polynomial at 1 yields

$$M(G_1(p, q)) = 2^{p+q+1} - p2^q - q2^p - 1.$$

Finally, to count the classes by size we again filter by the three cases (1), (2), and (3). In the first case, there are $2^p - 1$ ways for the edge $\{1, 2\}$ to be in an immorality with any of the p leaves at node 1, and there are $2^q - q$ possible patterns of immoralities that can occur among the q leaves at node 2. One of these $2^q - q$ patterns has class size $q + 1$ (the class with no immoralities), and all others have size one. Thus, case (1) yields $2^p - 1$ classes of size $q + 1$ and $(2^q - q - 1)(2^p - 1)$ classes of size 1. Similarly, case (2) yields $2^q - 1$ classes of size $p + 1$ and $(2^p - p - 1)(2^q - 1)$ classes of size 1. In case (3), if both sets of leaves contain no immoralities, then we get a single class of size $p + q + 2$. If the p leaves at node 1 contain at least one immorality, then all leaves at node 2 must be directed away from node 2, yielding $2^p - p - 1$ classes of size 1. Similarly, if the q leaves at node 2 contain at least one immorality, then we get another $2^q - q - 1$ classes of size one. Summing over these cases yields the desired formulae. $\square$

3. BOUNDING THE SIZE AND NUMBER OF MECs ON TREES

We begin this section by deriving upper and lower bounds on the number of MECs for trees on p nodes. We show that these bounds are achieved by the $(p-1)$ -star $G_1(p-1)$ and the p -path I_p , respectively. This result parallels the classic result of [37], which states that the number of independent sets in a tree on p nodes is bounded by the number of independent sets in $G_1(p-1)$ and I_p , respectively.

Theorem 3.1. *Let T_p be a tree on p nodes. Then*

$$F_{p-1} = M(I_p) \leq M(T_p) \leq M(G_1(p-1)) = 2^{p-1} - p + 1.$$

Proof. We first prove the upper bound on $M(T_p)$. Since T_p is a tree, it has precisely $p-1$ edges, and so there are 2^{p-1} edge orientations on T_p . Of these 2^{p-1} orientations, the p orientations given by selecting a unique source node in T_p all belong to the

same MEC. So there are at most $2^{p-1} - p + 1$ MECs for T_p . By Theorem 2.4, this bound is achieved by the $(p-1)$ -star $G_1(p-1)$.

To prove the lower bound, we use a simple inductive argument. Notice first that the bound is true when $p \leq 5$. Now recall that every tree on p nodes can be constructed in one of two ways: (1) attaching a leaf to a degree 1 node of a tree on $p-1$ nodes, or (2) attaching a leaf to a node of T_{p-1} that is a neighbor of a leaf. Thus, given a tree T_{p-1} on $p-1$ nodes, it suffices to show that when we construct T_p from T_{p-1} via (1) or (2), the number of MECs increases by at least F_{p-3} .

In case (1), we attach a leaf node v to a leaf u of T_{p-1} , whose only neighbor in T_{p-1} is some node w . The MECs on T_p then come in two types: either the edge $\{v, u\}$ is not in an immorality or it is in the immorality $v \rightarrow u \leftarrow w$. The number of classes in the first case is $M(T_{p-1})$ and the number of classes in the second case is $M(T_{p-1} \setminus u)$. So by the inductive hypothesis we have that

$$M(T_p) \geq M(T_{p-1}) + M(T_{p-1} \setminus u) \geq F_{p-2} + F_{p-3} = F_{p-1}.$$

In case (2), the leaf node v is attached to some node u of T_{p-1} that has at least one leaf w in T_{p-1} . The MECs on T_p contain two disjoint types of classes: classes in which the edge $\{v, u\}$ is not in an immorality and classes containing the immorality $v \rightarrow u \leftarrow w$. Similar to the previous case, it then follows from the inductive hypothesis that

$$M(T_p) \geq M(T_{p-1}) + M(T_{p-1} \setminus w) \geq F_{p-2} + F_{p-3} = F_{p-1},$$

which completes the proof. $\square$

We now derive bounds on the size of the MEC for a fixed DAG $\mathcal{T}_p$ on the underlying undirected graph T_p . These bounds will be computed in terms of the structure of the *essential graph* $\widehat{\mathcal{T}}_p$ of the MEC $[\mathcal{T}_p]$. Recall that the essential graph of an MEC $[\mathcal{G}]$ is a partially directed graph $\widehat{\mathcal{G}} := ([p], E, A)$, where the collection of arrows A in $\widehat{\mathcal{G}}$ are the arrows that point in the same direction for every member of the class, and the undirected edges E represent the arrows that change orientation to distinguish between members of the class; see [3]. The *chain components* of $\widehat{\mathcal{G}}$ are its undirected connected components, and its *essential components* are its directed connected components.

To see why it is reasonable to work with the essential graph to derive such bounds, consider the analysis of the MEC sizes for stars and bistars given in Theorems 2.4 and 2.5. In order to derive the possible sizes of these MECs, we implicitly counted all possible orientations of the undirected edges in the essential graph of each class. Since understanding the possible orientations of these edges is equivalent to knowing the size of the class, we will bound the size of the MEC of $\mathcal{T}_p$ in terms of the number and size of the chain components of $\widehat{\mathcal{T}}_p$. We will see that the computed bounds are tight, and that stars play an important role in achieving these bounds. We refer the reader to [3] for the basics relating to essential graphs.

In the following, we assume that the essential graph $\widehat{\mathcal{T}}_p$ has chain components $\tau_1, \tau_2, \dots, \tau_\ell$ for $\ell > 0$. We also assume that each τ_i is *nontrivial*; i.e. it has at least two vertices. We let $\mathcal{G}(\widehat{\mathcal{T}}_p)$ denote the directed subforest of the essential graph $\widehat{\mathcal{T}}_p$ consisting of all directed edges of $\widehat{\mathcal{T}}_p$, and we let $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m$ denote its connected components.

Lemma 3.2. *Let $\mathcal{T}_p$ be a directed tree on p nodes and $\widehat{\mathcal{T}}_p$ the corresponding essential graph. If $\widehat{\mathcal{T}}_p$ has chain components $\tau_1, \tau_2, \dots, \tau_\ell$, then the size of the Markov equivalence class $[\mathcal{T}_p]$ is*

$$\#[\mathcal{T}_p] = \prod_{i=1}^{\ell} |V(\tau_i)|.$$

Proof. Each element of $[\mathcal{T}_p]$ corresponds to one of the ways to direct the components $\tau_1, \dots, \tau_\ell$, each of which is a tree. Suppose we directed τ_i so that it has two source nodes s_1 and s_2 . Then along the unique path between s_1 and s_2 in the directed τ_i , there must lie an immorality that is not present in $\widehat{\mathcal{T}}_p$. Thus, the only admissible directions of the components τ_i have no more than one source node. Since every DAG has at least one source node, the number of admissible directions of each τ_i is precisely the number of ways to pick the unique source node of τ_i . This is precisely the number of vertices in τ_i , thereby completing the proof. $\square$

Theorem 3.3. *Let $\mathcal{T}_p$ be a directed tree on p nodes and $\widehat{\mathcal{T}}_p$ the corresponding essential graph. Suppose that $\widehat{\mathcal{T}}_p$ has $\ell > 0$ chain components $\tau_1, \tau_2, \dots, \tau_\ell$ and that the directed subforest $\mathcal{G}(\widehat{\mathcal{T}}_p)$ of $\widehat{\mathcal{T}}_p$ has $m \geq 0$ connected components $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m$. Then*

$$2^\ell \leq \#[\mathcal{T}_p] \leq \left(\frac{p-m}{\ell} \right)^\ell.$$

Proof. Notice first that the lower bound is immediate from Lemma 3.2 and the assumption that each τ_i is nontrivial. So it only remains to verify the proposed upper bound.

Let ℓ_i denote the number of chain components that are adjacent to ε_i for all $i \in [m]$. Since the chain components $\tau_1, \dots, \tau_\ell$ are all disjoint, it follows that

$$1 + \ell_i \leq |V(\varepsilon_i)|$$

for all $i \in [m]$. Therefore, a lower bound on the size of the number of nodes in the directed subforest $\mathcal{G}(\widehat{\mathcal{T}}_p)$ is given by

$$m + \sum_{i=1}^m \ell_i \leq |V(\mathcal{G}(\widehat{\mathcal{T}}_p))|.$$

A closed form for the sum $\sum_{i=1}^m \ell_i$ is recovered as follows. Consider a complete bipartite graph $K_{\ell, m}$ whose vertices are partitioned into two blocks A and B where $|A| = \ell$ and $|B| = m$. The possible ways to assemble the components $\tau_1, \dots, \tau_\ell$ and $\varepsilon_1, \dots, \varepsilon_m$ into an essential tree are in bijection with the spanning trees of $K_{\ell, m}$. For any such spanning tree T of $K_{\ell, m}$, each edge of T has exactly one vertex in each of A and B . Thus,

$$\sum_{i=1}^m \ell_i = \sum_{v \in A} \deg_T(v) = \sum_{v \in B} \deg_T(v).$$

Since T is a tree, it follows that

$$\sum_{i=1}^m \ell_i = \frac{\sum_{v \in A} \deg_T(v) + \sum_{v \in B} \deg_T(v)}{2} = \ell + m - 1. \quad (1)$$

Therefore,

$$2m + \ell - 1 \leq |V(\mathcal{G}(\widehat{\mathcal{T}}_p))|.$$

Moreover, since $\mathcal{T}_p$ has p vertices, and each edge of a spanning tree of $K_{\ell,m}$ corresponds to exactly one of the vertices shared by $\mathcal{G}(\widehat{\mathcal{T}}_p)$ and the chain components $\tau_1, \dots, \tau_\ell$, then we have that

$$\sum_{j=1}^{\ell} |V(\tau_j)| = p + m + \ell - 1 - |V(\mathcal{G}(\widehat{\mathcal{T}}_p))|. \quad (2)$$

Now by Lemma 3.2 and the arithmetic-geometric mean inequality, we have

$$\#[\mathcal{T}_p] = \prod_{j=1}^{\ell} |V(\tau_j)| \leq \left(\frac{\sum_{j=1}^{\ell} |V(\tau_j)|}{\ell} \right)^{\ell}.$$

Thus, by applying equation 2, we conclude that

$$\#[\mathcal{T}_p] \leq \left(\frac{p + m + \ell - 1 - |V(\mathcal{G}(\widehat{\mathcal{T}}_p))|}{\ell} \right)^{\ell} \leq \left(\frac{p + m + \ell - 1 - (2m + \ell - 1)}{\ell} \right)^{\ell},$$

and so $\#[\mathcal{T}_p] \leq ((p - m)/\ell)^{\ell}$, which completes the proof. $\square$

We now examine the tightness of the bounds in Theorem 3.3 by considering some special cases. Notice first that the lower bound is tight exactly when each chain component is a single edge. The upper bound is tight exactly when $|V(\mathcal{G}(\widehat{\mathcal{T}}_p))| = 2m + \ell - 1$ and each chain component has exactly $\frac{p-m}{\ell}$ vertices.

Corollary 3.4. *Suppose $\mathcal{G}(\widehat{\mathcal{T}}_p)$ has precisely one connected component, i.e., $\mathcal{G}(\widehat{\mathcal{T}}_p)$ is a directed tree. Then*

$$2^{\ell} \leq \#[\mathcal{T}_p] \leq \left(\frac{p-1}{\ell} \right)^{\ell},$$

and every directed tree $\mathcal{T}_p$ for which the upper bound is tight has the same subtree $\mathcal{G}(\widehat{\mathcal{T}}_p)$, namely $G_1(\ell)$ with all edges directed inwards.

Proof. The statement of the bounds is immediate from Theorem 3.3. So we only need to verify the claim on the tightness of the upper bound. It follows from the more general bounds described above, that the upper bound is tight exactly when $|V(\mathcal{G}(\widehat{\mathcal{T}}_p))| = \ell + 1$ and each chain component has exactly $\frac{p-1}{\ell}$ vertices. Since the chain components $\tau_1, \dots, \tau_\ell$ are all distinct and $\mathcal{G}(\widehat{\mathcal{T}}_p)$ is a directed tree with $\ell + 1$ vertices, then each τ_j is adjacent to exactly one of the ℓ vertices of $\mathcal{G}(\widehat{\mathcal{T}}_p)$, and there remains only one vertex to connect these ℓ vertices. Therefore, the skeleton of $\mathcal{G}(\widehat{\mathcal{T}}_p)$ is the star $G_1(k+1)$. Moreover, since all essential edges in $\widehat{\mathcal{T}}_p$ are exactly the edges of $\mathcal{G}(\widehat{\mathcal{T}}_p)$, then all edges of $\mathcal{G}(\widehat{\mathcal{T}}_p)$ must be directed inwards towards the center node. An example of a graph for which this upper bound is tight is presented on the left in Figure 4. $\square$

Corollary 3.5. *Suppose $\widehat{\mathcal{T}}_p$ has precisely one chain component τ_1 . Then*

$$m \leq \#[\mathcal{T}_p] \leq p - 2m,$$

and both bounds are tight when $\tau_1 = G_1(m-1)$.

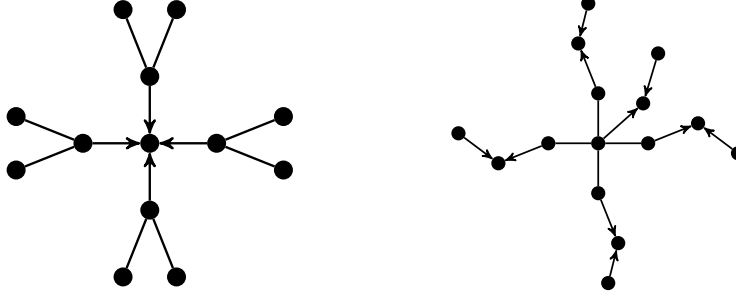


FIGURE 4. Graphs for which the bounds in Corollary 3.5 are tight when $m = 1$ (left) and when $k = 1$ (right).

Proof. By Lemma 3.2 we know that $\#[\mathcal{T}_p] = |V(\tau_1)|$; so the bounds presented here are bounds on the size of the vertex set of the chain component τ_1 . Since the connected components $\varepsilon_1, \dots, \varepsilon_m$ of $\mathcal{G}(\widehat{\mathcal{T}}_p)$ are all disjoint, we know that τ_1 contains at least m vertices. On the other hand, since each ε_i contains at least one immorality and attaches to τ_1 at precisely one node, then each ε_i contains at least two nodes that are not also nodes of τ_i . A graph for which the bounds are simultaneously tight is depicted on the right in Figure 4. Notice that the chain component τ_1 is $G_1(m - 1)$. $\square$

Corollary 3.4 and Corollary 3.5 suggest the important role of the maximum degree of a graph for the size of MECs. This is further supported and discussed via the results in the next section and the simulations in Section 5.2.

4. CLASSIC FAMILIES OF TREES

In this section, we study some classic families of trees that arise naturally in both, applied and theoretical contexts. Namely, we will study the graph polynomials $M(G; x)$ for *spider graphs*, *caterpillar graphs*, and *complete binary trees*. A *spider graph* (or *star-like tree*) is any tree containing precisely one node with degree greater than two, a *caterpillar graph* is any tree for which deleting all leaves results in a path, and a complete binary tree is a tree for which every nonleaf node (except for possibly a root node) has precisely three neighbors. Caterpillars and complete binary trees play important roles for modeling events in time, as for example in phylogenetics. Caterpillars and spiders also provide large families of supporting examples for long-standing conjectures about well-studied generating functions associated to trees. Alavi, Maldi, Schwenk, and Erdős conjectured that the independence polynomial of every tree is unimodal [2], and Stanley conjectured that the chromatic symmetric function is a complete graph isomorphism invariant for trees [44]. In [27, 28] and [32] the authors, respectively, verify that these conjectures hold for caterpillars and (some) spiders. We show in the following that these important families of graphs also yield nice properties for the generating polynomial $M(G; x)$.

In Section 4.1 we provide a formula for $M(G; x)$ for spider graphs that generalizes our formula for stars and paths given in Section 2. Using these formulae we compute expressions for $M(G)$ that extend classical identities of the Fibonacci numbers. The methods for computing $M(G; x)$ for spiders generalizes to a multivariate formula for

$M(G; x)$ for arbitrary trees with interesting combinatorial structure, which will also be described. In Section 4.2 we recursively compute $M(G; x)$ for the caterpillars. Using this recursive formula, we observe that these polynomials are all unimodal and estimate the expected number of immoralities in a randomly selected MEC on a caterpillar. Finally, in Section 4.3 we compute the number of MECs for a complete binary tree, and study the rate at which this value increases.

4.1. Spiders. We call the unique node of degree more than two in a spider its *center* node. A spider G on n nodes with center node of degree k corresponds to a partition $\lambda = (\lambda_1, \dots, \lambda_k)$ of $n - 1$ into k parts. Following the standard notation, we assume $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k > 0$. Here, λ_i denotes the number of vertices on the i^{th} leg of G ; i.e., the i^{th} maximal connected subgraph of G in which every vertex has degree at most two. Conversely, given a partition λ of $n - 1$ into k parts, we write G_λ for the corresponding spider graph.

In the following, we label the vertices of G_λ such that p_0 denotes the center node and p_{ij} , for $1 \leq i \leq k$ and $1 \leq j \leq \lambda_i$, denotes the j^{th} node from p_0 along the i^{th} leg of G_λ . For a subset $S \subset [k]$, define the following polynomial:

$$L(S; x) := \left(\prod_{i \in S} M(I_{\lambda_i - 1}; x) \right) \left(\prod_{i \in [k] \setminus S} M(I_{\lambda_i}; x) \right).$$

We then have the following formula for the generating polynomial $M(G_\lambda; x)$.

Theorem 4.1. *Let G_λ denote the spider on n nodes with center node of degree k and partition λ of $n - 1$ into k parts. If λ has ℓ parts of size one, then*

$$M(G_\lambda; x) = \sum_{j=0}^{k-\ell} \left(\sum_{S \in \binom{[k-\ell]}{j}} L(S; x) \right) x^j M(G_1(k-j); x).$$

Proof. To arrive at this formula, simply notice that all possible placements of immoralities can be computed as follows: First choose a subset of the $k - \ell$ nodes $\{p_{ij} : \lambda_i > 1, i \in [k]\}$ at which to place immoralities. Call this set S . Since the nodes in $\{p_{ij} : \lambda_i > 1, i \in [k]\} \setminus S$ are not immoralities then all remaining immoralities are either at the center node p_0 , which are counted by $M(G_1(k - |S|); x)$, or they are further down the legs of the spider, which are counted by $L(S; x)$. $\square$

The general formula in Theorem 4.1 specializes to $M(G_1(p-1); x)$ when $\lambda = (1, 1, \dots, 1)$ is the partition of $p - 1$ into $p - 1$ parts; i.e., when $G_\lambda = G_1(p - 1)$. Similarly, for $k = 2$, it reduces to $M(I_p; x)$. It also yields a nice formula for the number of MECs on the spiders with $\lambda = (m, m, \dots, m)$ a partition of mk into k parts.

Corollary 4.2. *For $k > 1$ and $m \geq 1$, the spider G_λ on $mk + 1$ nodes with partition $\lambda = (m, m, \dots, m)$ of mk into k parts has*

$$M(G_\lambda) = F_{m+1}^k - k F_{m-1} F_m^{k-1}.$$

Proof. For $m = 1$ we have that $G_\lambda = G_1(k)$, and the above formula reduces to $2^k - k = M(G_1(k))$. For $k > 1$, we simplify the formula given in Theorem 4.1 to

$$M(G_\lambda; x) = \sum_{j=0}^k \binom{k}{j} (x M(I_{m-1}; x))^j M(I_m; x)^{k-j} M(G_1(k-j); x).$$

Evaluating at $x = 1$ yields

$$\begin{aligned}
M(G_\lambda) &= \sum_{j=0}^k \binom{k}{j} F_{m-2}^j F_{m-1}^{k-j} (2^{k-j} - (k-j)), \\
&= \sum_{j=0}^k \binom{k}{j} F_{m-2}^j (2F_{m-1})^{k-j} - \sum_{j=0}^k \binom{k}{j} (k-j) F_{m-2}^j (F_{m-1})^{k-j}, \\
&= (F_{m-2} + 2F_{m-1})^{k-j} - kF_{m-1} (F_{m-2} + F_{m-1})^{k-1}, \\
&= F_{m+1}^k - kF_{m-1} F_m^{k-1},
\end{aligned}$$

which completes the proof. $\square$

Remark 4.1. In the special case of Corollary 4.2 for which $k = 2$ we have that $G_\lambda = I_{2m+1}$, and so $M(G_\lambda) = F_{2m}$ by Theorem 2.1. In Corollary 4.2, we see that the formula for $M(G_\lambda)$ given by Theorem 4.1 is computing the Fibonacci number F_{2m} via a classic identity discovered by Lucas in 1876 (see for instance [26]):

$$F_{2m} = F_{m+1}^2 - 2F_{m-1}F_m = F_m^2 + F_{m-1}^2.$$

Notice that the same expression does not hold for the generating polynomials:

$$M(G_\lambda; x) \neq M(I_{m+2}; x)^k - kM(I_m; x)M(I_{m+1}; x)^{k-1}.$$

This is because $M(G_1(p); x) = 1 + \sum_{k \geq 2} \binom{p}{k} x^{\binom{k}{2}}$ as opposed to $1 + \sum_{k \geq 2} \binom{p}{k} x^k$. However, when the formula for $M(G_\lambda; x)$ used in the proof of Corollary 4.2 is evaluated at $x = 1$, the exponents in the formula for $M(G_1(p); x)$ become irrelevant. For instance, in the case when $\lambda = (2, 2)$, we have that

$$\begin{aligned}
M(G_\lambda; x) &= x^2 + 3x + 1, \text{ but} \\
M(I_{m+2}; x)^k - kM(I_m; x)M(I_{m+1}; x)^{k-1} &= 4x^2 + 2x - 1.
\end{aligned}$$

However, evaluating both polynomials at $x = 1$ results in the Fibonacci number $F_4 = 5$, as predicted by Corollary 4.2.

We end this section with a remark and example illustrating the more general consequences of the techniques used in the computation of $M(G_\lambda; x)$ in Theorem 4.1.

Remark 4.2. It is natural to ask if the recursive approach used to prove Theorem 4.1 generalizes to arbitrary trees. In particular, it would be nice if for any tree T , the polynomial $M(T; x)$ can be expressed as

$$M(T; x) = \sum_{\alpha = (\alpha_2, \dots, \alpha_{n-1}) \in \mathbb{Z}_{\geq 0}^{n-2}} c_\alpha \mathbf{s}^\alpha, \quad (3)$$

where $s_i := M(G_1(i); x)$ for $i = 2, \dots, n-1$, $\mathbf{s}^\alpha := s_2^{\alpha_2} s_3^{\alpha_3} \cdots s_{n-1}^{\alpha_{n-1}}$, and the c_α are polynomials in x with nonnegative integer coefficients. On the one hand, there exists an (albeit cumbersome) recursion for computing $M(T; x)$ that generalizes the one used in Theorem 4.1. On the other hand, this recursion will not yield an expression of the form in equation (3) unless it has at most one node with degree more than two. Instead, if we take

$$a_p := \sum_{k \geq 2} \binom{p-1}{k} x^{\binom{k}{2}} \quad \text{and} \quad b_p := \sum_{k \geq 2} \binom{p-1}{k-1} x^{\binom{k}{2}},$$

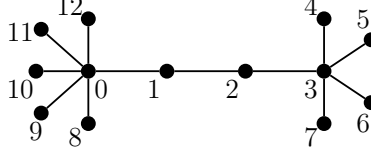


FIGURE 5. The tree for Example 4.1.

then we can express $M(T; x)$ as

$$M(T; x) = \sum_{\alpha=(\alpha_1, \alpha_2, \alpha_3) \in \mathbb{Z}_{\geq 0}^{n-2} \times \mathbb{Z}_{\geq 0}^{n-2} \times \mathbb{Z}_{\geq 0}^{n-2}} c_{\alpha} \mathbf{s}^{\alpha_1} \mathbf{a}^{\alpha_2} \mathbf{b}^{\alpha_3}, \quad (4)$$

where $\mathbf{a}^{\alpha}$ and $\mathbf{b}^{\alpha}$ are defined analogously to $\mathbf{s}^{\alpha}$, and the c_{α} are polynomials in x with nonnegative integer coefficients. The algorithm resulting in the expression for $M(T; x)$ given in equation (4) is the intuitive generalization of Theorem 4.1. Since it is technical to formalize, we here only illustrate it with Example 4.1.

Example 4.1. Consider the tree T on 12 nodes depicted in Figure 5. We follow the same approach for counting MECs in T that we used to count the MECs in G_{λ} in Theorem 4.1. That is, we select a center node, choose a collection of immoralities at its nonleaf neighbors, and count the possible classes containing these immoralities. Thinking of node 0 as the analogous vertex to the center node of a spider, we notice that it has precisely one nonleaf neighbor, namely node 1. The MECs on T with node 1 in an immorality are counted by xs_5^2 . Now consider those MECs on T for which 1 is not in an immorality. Analogous to the proof of Theorem 4.1, we must consider the MECs on the 6-star with center node 0 and leaves 1, 8, 9, 10, 11, and 12. Notice b_6 enumerates the MECs on this 6-star that use the arrow $0 \leftarrow 1$, and $a_6 + 1$ enumerates those MECs not using this arrow. For those enumerated by b_6 , we then count the number of MECs on the induced subtree T' with vertex set [7]. This gives $b_6 M(T'; x)$.

For the MECs enumerated by $a_6 + 1$, we must consider more carefully the structure of immoralities on T' . The constant 1 counts the choice of no immoralities on the 6-star, and this yields $1M(T'; x)$ MECs on T . On the other hand, a_6 counts those classes on the 6-star with at least one immorality using the arrow $0 \leftarrow 1$. For these, we take node 2 as the center node of T' , which has precisely one non-leaf neighbor, node 3. The ways in which node 3 can be in an immorality are counted by b_5 . If node 3 is not in an immorality, then either 2 is in an immorality or there are no immoralities on T' . This yields $a_6(1 + b_5 + xs_4)$. Using the same techniques, we compute that $M(T'; x) = s_2s_4 + b_5$. Combining these formulae yields

$$M(T; x) = xs_5^2 + s_2s_4 + s_2s_4b_6 + b_5 + b_5b_6 + xs_4a_6 + a_6 + a_6b_5.$$

In general, this iterative process of picking a center node for a tree T , choosing immorality placements for its nonleaf neighbors, and then enumerating the resulting possible MECs based on these choices results in an expression of the form given by equation (4). The monomial $\mathbf{s}^{\alpha_1} \mathbf{a}^{\alpha_2} \mathbf{b}^{\alpha_3}$ enumerates the possible placements of immoralities at the chosen sequence of center nodes and the coefficient polynomial c_{α} is enumerating the ways to fix immoralities at their nonleaf neighbors to allow for these placements. $\square$

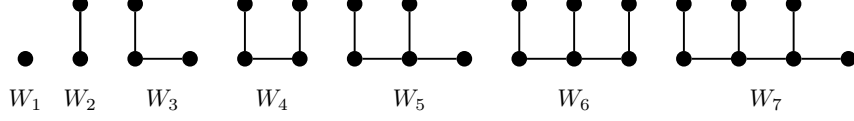


FIGURE 6. The first few caterpillar graphs.

Theorem 4.1 demonstrates that for some trees the expression for $M(T; x)$ given by the algorithmic approach described in Example 4.1 can have nice coefficient polynomials c_α . It is important to notice that the expression of $M(T; x)$ given in equation (4) is dependent of the initial choice of center node. However, as exhibited by Theorem 4.1, a well-chosen initial center node and number of iterations of this decomposition can yield nice combinatorial expressions for $M(T; x)$ of the form (3) and/or (4). For example, if T is the spider graph, one iteration of this decomposition initialized at the spider's center yields coefficient polynomials c_α that are products of Fibonacci polynomials, and when all legs are the same length, they are therefore real-rooted, log-concave, and unimodal. It would be interesting to know whether other families of trees yield coefficient polynomials c_α with nice combinatorial properties. Moreover, it is unclear if for every tree T the polynomial $M(T; x)$ admits an expression as in equation (3).

4.2. Caterpillars. We denote the caterpillar graph W_p as

$$W_p := \begin{cases} G_{\frac{p}{2}}(1, 1, \dots, 1) & \text{if } p \text{ is even,} \\ G_{\frac{p+1}{2}}(1, 1, \dots, 1, 0) & \text{if } p \text{ is odd.} \end{cases}$$

The first few caterpillar graphs are depicted in Figure 6. Since the caterpillar graphs are closely related to paths, we would expect that a similar recursive approach also works for counting the number of MECs on W_p . Indeed, with the following theorem, we provide a recursive formula for $M(W_p; x)$.

Theorem 4.3. *Let $\mathbb{W}_p := M(W_p; x)$ for $p \geq 1$. These generating polynomials satisfy the recursion with initial conditions*

$$\mathbb{W}_1 = 1, \quad \mathbb{W}_2 = 1, \quad \mathbb{W}_3 = 1 + x, \quad \mathbb{W}_4 = 1 + 2x,$$

and for $p \geq 5$

$$\mathbb{W}_p = \begin{cases} \mathbb{W}_{p-1} + x\mathbb{W}_{p-2} & \text{for } p \text{ odd,} \\ (x+2)\mathbb{W}_{p-2} + (x^3 - x^2 + x - 2)\mathbb{W}_{p-3} + (x^2 + 1)\mathbb{W}_{p-4} & \text{for } p \text{ even.} \end{cases}$$

Proof. Notice first that when p is even, we can simply apply the Fibonacci recursion

$$M(G_{\frac{p}{2}}(1, 1, \dots, 1); x) = M(G_{\frac{p}{2}}(1, 1, \dots, 0)) + xM(G_{\frac{p}{2}-1}(1, 1, \dots, 1); x).$$

The recursion is based on whether or not the final edge is contained within an immorality.

Now let $p = 2k + 1$ be odd. We first show that

$$\mathbb{W}_p = \mathbb{W}_{p-1} + (x^3 + x)\mathbb{W}_{p-3} + x\mathbb{W}_{p-2} - x^2 \sum_{j=2}^{\lfloor \frac{p}{2} \rfloor} \mathbb{W}_{p-2j-1}.$$

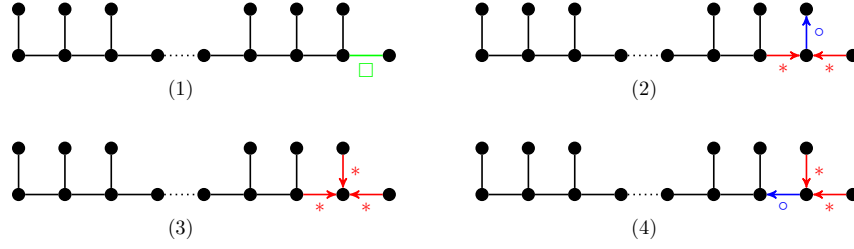


FIGURE 7. The four cases for the recursion on the caterpillar graph for p odd.

This recursion can be detected by considering the ways in which the final edge can or cannot be in an immorality. That is, either it is not in an immorality, or it is in an immorality with some nonempty subset of edges adjacent to it, as depicted in Figure 7. Collectively, cases (1), (2), and (3) yield

$$\mathbb{W}_{p-1} + (x^3 + x)\mathbb{W}_{p-3}$$

MECs. On the other hand, case (4) yields $x\mathbb{W}_{p-2}$ minus some over-counted cases. The over-counted cases correspond to exactly when the first immorality to the right of the one depicted in case (4) points towards the right, as depicted in Figure 8. Each such case would naturally force one more unspecified immorality. Thus, the total number of MECs counted by case (4) is

$$x\mathbb{W}_{p-2} - x^2 \sum_{j=2}^{\lfloor \frac{p}{2} \rfloor} \mathbb{W}_{p-2j-1}.$$

Since $p-1$ is even, we may apply the Fibonacci recursion to $\mathbb{W}_{p-1}$ to obtain

$$\mathbb{W}_p - (x+1)\mathbb{W}_{p-2} = (x^3 + 2x)\mathbb{W}_{p-3} - x^2 \sum_{j=2}^{\lfloor \frac{p}{2} \rfloor} \mathbb{W}_{p-2j-1}.$$

We then consider the difference between $\mathbb{W}_p - (x+1)\mathbb{W}_{p-2}$ and $\mathbb{W}_{p-2} - (x+1)\mathbb{W}_{p-4}$, and repeatedly apply the Fibonacci recursion to the even terms. The result is

$$\mathbb{W}_p - (x+2)\mathbb{W}_{p-2} = (x^3 + x - 1)\mathbb{W}_{p-4} + (x^3 - x^2 + x - 2)(\mathbb{W}_{p-3} - \mathbb{W}_{p-4}).$$

This simplifies to

$$\mathbb{W}_p = (x+2)\mathbb{W}_{p-2} + (x^3 - x^2 + x - 2)\mathbb{W}_{p-3} + (x^2 + 1)\mathbb{W}_{p-4},$$

thereby completing the proof. $\square$

The first few polynomials $M(W_p; x)$ for $1 \leq p \leq 14$, and the number of MECs on W_p , are displayed in Table 1. These polynomials all appear to be unimodal.



FIGURE 8. The over-counted cases of case (4) in the caterpillar recursion for p odd.

$M(W_p)$	$M(W_p; x)$
1	1
1	1
2	$x + 1$
3	$2x + 1$
7	$x^3 + x^2 + 4x + 1$
10	$x^3 + 3x^2 + 5x + 1$
22	$3x^4 + 3x^3 + 8x^2 + 7x + 1$
32	$4x^4 + 6x^3 + 13x^2 + 8x + 1$
70	$x^6 + 6x^5 + 13x^4 + 16x^3 + 23x^2 + 10x + 1$
102	$x^6 + 10x^5 + 19x^4 + 29x^3 + 31x^2 + 11x + 1$
222	$5x^7 + 13x^6 + 39x^5 + 46x^4 + 59x^3 + 46x^2 + 13x + 1$
324	$6x^7 + 23x^6 + 58x^5 + 75x^4 + 90x^3 + 57x^2 + 14x + 1$
704	$x^9 + 15x^8 + 39x^7 + 97x^6 + 147x^5 + 158x^4 + 153x^3 + 77x^2 + 16x + 1$
1028	$x^9 + 21x^8 + 62x^7 + 155x^6 + 222x^5 + 248x^4 + 210x^3 + 91x^2 + 17x + 1$

TABLE 1. The number of MECs of W_p for $1 \leq p \leq 14$, and the associated polynomial generating functions $M(W_p; x)$.

Using the recursion in Theorem 4.3 we can estimate that the immorality number of W_p is $m(W_p) = \lfloor \frac{p}{2} \rfloor + \lfloor \frac{p}{4} \rfloor$, and that the expected number of immoralities in a randomly chosen MEC on W_p approaches $\left\lfloor \frac{m(W_p)}{2} \right\rfloor$. As an immediate corollary to Theorem 4.3, we get a recursion for the number of MECs $M(W_p)$.

Corollary 4.4. *The number of MECs for the caterpillar graph W_p is given by the recursion*

$$M(W_1) = 1, \quad M(W_2) = 1, \quad M(W_3) = 2, \quad M(W_4) = 3,$$

and for $p \geq 5$

$$M(W_p) = \begin{cases} M(W_{p-1}) + M(W_{p-2}) & \text{if } p \text{ is even,} \\ 3M(W_{p-2}) + M(W_{p-4}) - M(W_{p-5}) & \text{if } p \text{ is odd.} \end{cases}$$

4.3. Complete Binary Trees. In the following, we let T_k denote the complete binary tree containing $2^k - 1$ nodes and A_k denote the additive tree constructed by adding one leaf to the root node of T_k . These two trees are depicted in Figure 9 for $k = 3$.

We will now use a series of recursions to enumerate the number of MECs on T_k and A_k . We will then show that the ratio $\frac{M(A_k)}{M(T_k)} < 4$, which means that adding

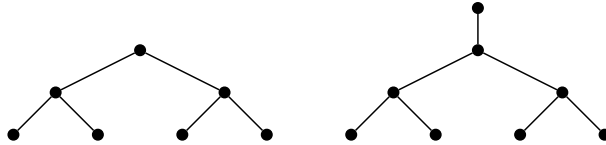
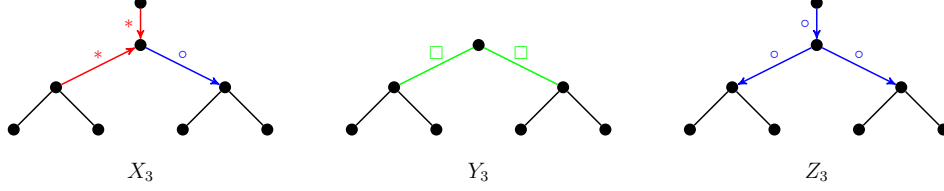


FIGURE 9. The complete binary tree T_3 is depicted on the left and the additive tree A_3 is depicted on the right.

FIGURE 10. From left-to-right, the graphs X_3 , Y_3 , and Z_3 .

an edge to the root of a complete binary tree increases the number of MECs by at most a factor of 4. In practice, we observed that the factor is around 2 for large k .

Before providing a recursion for $M(T_k)$ and $M(A_k)$, we introduce three new graph structures X_k , Y_k , and Z_k in order to help simplify our recursions. Similar to Section 2.2, in the following it will be helpful to label edges that have specified roles in certain MECs. The green edges (also labeled with $\square$) indicate that these edges cannot be involved in any immorality. The red arrows (also labeled with $*$) indicate a fixed immorality in the partially directed graph, and the blue arrows (also labeled with $\circ$) represent fixed arrows that are not in immoralities.

- (1) Let X_k denote the partially directed tree whose skeleton is A_k and for which there is exactly one immorality at the child of the root (note that the root of A_k has degree 1).
- (2) Let Y_k denote the number of MECs on a complete binary tree with $2^k - 1$ nodes such that the root's edges are not involved in any immoralities.
- (3) Let Z_k denote the number of MECs on an additive tree with 2^k nodes such that there are edges directed from the root r to its child c and from c to each of its children.

The graphs X_3 , Y_3 , and Z_3 are depicted from left-to-right in Figure 10. Now we have the following series of recursions for the graphs listed above.

Theorem 4.5. *The following recursions hold for the partially directed graphs T_k , A_k , X_k , Y_k , and Z_k :*

- (a) $M(T_k) = M(A_{k-1})^2 + M(Y_k)$ with $M(T_1) = 1$,
- (b) $M(A_k) = M(T_k) + 2M(X_k) + M(T_{k-1})^2$ with $M(A_1) = 1$,
- (c) $M(X_k) = M(T_{k-1})\sqrt{M(Z_k)}$ with $M(X_1) = 1$,
- (d) $M(Y_k) = 2M(Z_{k-1})M(T_{k-1}) - M(Z_{k-1})^2$ with $Y_1 = 1$, and
- (e) $M(Z_k) = (2M(X_{k-1}) + M(T_{k-2})^2 + M(Z_{k-1}))^2$ with $Z_1 = Z_2 = 1$.

We first prove statements (e), (c), (d) in this order and then use them to prove statements (b) and (a).

Proof of statement (e). We prove this by analyzing the cases on the left subgraph of Z_k and consider possible immoralities at node s in Figure 11.

- (1) If node s has exactly one immorality (as in the leftmost figure), then this substructure contributes exactly $M(X_{k-1})$ MECs. By symmetry, there are two ways in which node s can have exactly one immorality, which means these cases contribute $2M(X_{k-1})$ MECs.
- (2) If node s has three immoralities (as in the center figure), then this substructure contributes exactly $M(T_{k-2})^2$ MECs as we may treat nodes u, v as roots of complete binary trees T_{k-2} .

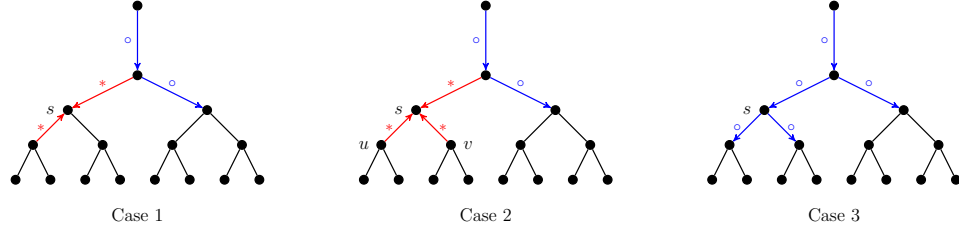


FIGURE 11.

- (3) If node s has no immoralities (as in the rightmost figure), then this substructure contributes exactly $M(Z_{k-1})$ MECs as we may treat the left subgraph as the graph Z_{k-1} .

Finally, as we have just considered the cases on the left subgraph of Z_k and as the immoralities on the right subgraph of Z_k are independent of the immoralities on the left subgraph, we square the number of MECs on the left subgraph to conclude that $M(Z_k) = (2M(X_{k-1}) + M(T_{k-2})^2 + M(Z_{k-1}))^2$.

Proof of statement (c). Suppose we label two nodes p and q in X_k as in Figure 12. By treating node p as the root of the complete binary tree, and by treating node q as node s in the proof of statement (e), we directly have that $M(X_k) = M(T_{k-1})\sqrt{M(Z_k)}$.

Proof of statement (d). We will prove the desired recursion by considering the equivalence classes for which the edges e_a and e_b in Figure 13 are directed towards the root or away from the root.

- (1) Suppose that edge e_a is directed away from the root, then edge e_b can always be directed so that it is not in an immorality at the root's right child. Thus we can consider the root's right child to be the root of the complete binary tree T_{k-1} . Now since there cannot be an immorality at the root's left child, the left subgraph of the root can be treated as the root of the subgraph Z_{k-1} . This case thus gives us $M(Z_{k-1})M(T_{k-1})$ MECs.
- (2) Suppose that edge e_b is now directed away from the root, then this case is symmetric to the case above and so there are again $M(Z_{k-1})M(T_{k-1})$ MECs formed.
- (3) In the above cases we have double-counted the cases where the edges e_a and e_b are both directed away from the root. Thus we must subtract the

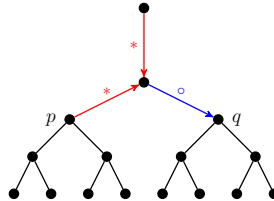


FIGURE 12.

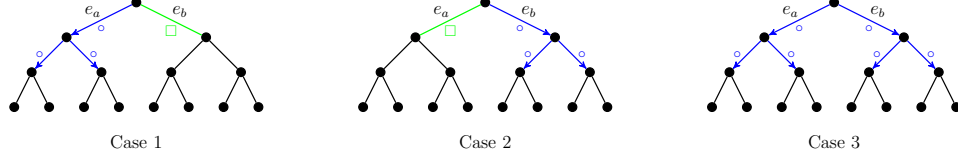


FIGURE 13.

number of MECs formed in this case. However, in this case the left and right subgraphs from the root both represent Z_{k-1} . Thus, there are $M(Z_{k-1})^2$ MECs in this case.

Hence we have that $M(Y_k) = 2M(Z_{k-1})M(T_{k-1}) - M(Z_{k-1})^2$.

Proof of statement (b). To prove recursion (b), we will consider the three possible cases of immoralities that can occur at the child c of the root as depicted in Figure 14.

- (1) In the leftmost figure, if there is no immorality formed by the edge from the root to c , then c can be treated as the root of the complete binary tree T_k . This case contributes $M(T_k)$ MECs.
- (2) In the center figure, if there is exactly one immorality formed by the edge from the root to s , then the root can be treated as the root of the tree X_k . This case contributes $2M(X_k)$ MECs, as there are two ways in which the edge from the root to c can be in exactly one immorality.
- (3) In the rightmost figure, if there are three immoralities formed by the edge from the root to c , then the children of c can be treated as roots of complete binary trees T_{k-1} . This case contributes $M(T_{k-1})^2$ MECs.

Thus, summing over the three cases we have that $M(A_k) = M(T_k) + 2M(X_k) + M(T_{k-1})^2$.

Proof of statement (a). We can consider the following four cases depicted in Figure 15 based on the immoralities formed by the root's edges e_a and e_b .

- (1) If the edges e_a and e_b form an immorality at the root, then the root's children p and q can be treated as roots of complete binary trees T_{k-1} . This case contributes $M(T_{k-1})^2$ MECs.
- (2) If the edge e_a forms at least one immorality at p but edge e_b is not in any immoralities, then edge q can be treated as the root of a complete binary tree T_{k-1} . Now p can have exactly one immorality, in which case the left

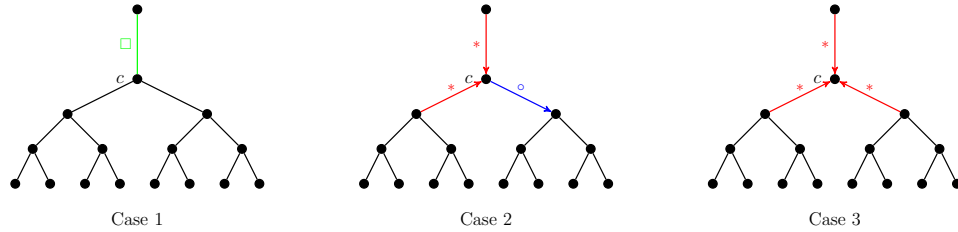


FIGURE 14.

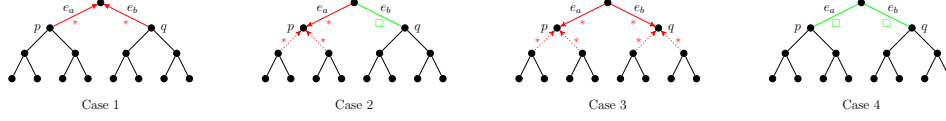


FIGURE 15.

subgraph of the root is the structure X_{k-1} or p can have three immoralities, in which case the children of p can each be treated as the root of a complete binary tree T_{k-2} . Now by symmetry we may consider immoralities formed by the edge e_b as well, which will double the number of MECs formed. Thus, there are $2M(T_{k-1})[2M(X_{k-1}) + M(T_{k-1})^2]$ MECs.

- (3) If the edges e_a and e_b form immoralities at p and q , then by following the reasoning in the previous case, there are $2M(X_{k-1}) + M(T_{k-2})^2$ MECs formed.
- (4) If the edges e_a and e_b form no immoralities, then the remaining graph is simply the structure Y_k . This case contributes $M(Y_k)$ MECs.

Summing over the different cases we have that

$$\begin{aligned}
 M(T_k) &= M(T_{k-1})^2 + 2M(T_{k-1})[2M(X_{k-1}) + M(T_{k-1})^2] + 2M(X_{k-1}) \\
 &\quad + M(T_{k-2})^2 + M(Y_k), \\
 &= [M(T_{k-1}) + 2M(X_{k-1}) + M(T_{k-2})^2]^2 + M(Y_k), \\
 &= M(A_{k-1})^2 + M(Y_k).
 \end{aligned}$$

This completes the proof of Theorem 4.5. $\square$

Now that we have recursions for T_k and A_k , we can establish a bound on the number of MECs given by adding an edge to the root of T_k to produce A_k . In order to do this, we will use the following lemma.

Lemma 4.6. *For the partially directed graphs T_k and Z_k we have that*

$$M(Z_k) < M(T_k).$$

Proof. If we omit the root and its edge from the graph Z_k , then we see that every MEC formed in Z_k can also be formed in T_k . Further, since the MEC in T_k with an immorality at the root cannot appear in Z_k , we have a strict inequality. Hence, we have that $M(Z_k) < M(T_k)$. $\square$

Now we show that adding an edge to the root of T_k increases the number of MECs by at most 4.

Theorem 4.7. *The number of MECs on A_k and T_k satisfy*

$$1 < \frac{M(A_k)}{M(T_k)} < 4.$$

Proof. First we let $R_k = \frac{M(A_k)}{M(T_k)}$ and $S_{k-1} = \frac{M(T_k)}{M(T_{k-1})^2}$. By equation (b) of Theorem 4.5 we know that

$$M(A_k) = M(T_k) + 2M(X_k) + M(T_{k-1})^2,$$

and hence by equation (c) of Theorem 4.5

$$\begin{aligned} R_k &= 1 + \frac{2M(X_k)}{M(T_k)} + \frac{M(T_{k-1})^2}{M(T_k)}, \\ &= 1 + \frac{2M(T_{k-1})\sqrt{M(Z_k)}}{M(T_k)} + \frac{M(T_{k-1})^2}{M(T_k)}. \end{aligned}$$

Thus, it follows by Lemma 4.6 that

$$R_k < 1 + \frac{2}{\sqrt{S_{k-1}}} + \frac{1}{S_{k-1}}$$

and hence

$$R_k < \left(1 + \frac{1}{\sqrt{S_{k-1}}}\right)^2 < \left(1 + \frac{1}{\sqrt{1}}\right)^2,$$

which completes the proof. $\square$

5. BEYOND TREES: OBSERVATIONS FOR TRIANGLE FREE GRAPHS

We end this paper with an analysis of the natural generalization of trees, the triangle-free graphs. As we will see, much of the intuition for the distribution of immoralities and number of MECs on trees carries over into the more general context of triangle-free graphs. However explicitly computing the generating functions $M(G; x)$ and $S(G; x)$ becomes increasingly difficult. In Section 5.1, we illustrate the increasing level of difficulty in computing these generating functions for triangle-free, non-tree, graphs by computing $M(G; x)$ and $S(G; x)$ for the complete bipartite graph $K_{2,p}$. In Section 5.2, we then take a computational approach to this problem, and we study the number and size of MECs relative to properties of the skeleton. Using data collected by a program described in [39], we examine the number and size of MECs on all connected graphs for $p \leq 10$ nodes and all triangle-free graphs for $p \leq 12$ nodes. We compare the number of MECs and their sizes to skeletal properties including average degree, maximum degree, clustering coefficient, and the ratio of the number of immoralities in the MEC to the number of induced 3-paths in the skeleton. For triangle-free graphs, we see that much of the intuition captured by the results of the previous sections extend into this setting. In particular, the number and distribution of high degree nodes in a triangle-free skeleton plays a key role in the number and sizes of MECs. Finally, unlike $S(G; x)$, we can see using graphs on few nodes that the polynomial $M(G; x)$ is not a complete graph isomorphism invariant for connected graphs on p nodes. For instance, the two graphs on four nodes in Figure 16 both have $M(G; x) = 1 + 2x + x^2$. However, using this program, we verify that $M(G; x)$ is a complete graph isomorphism

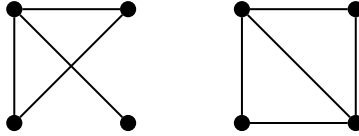


FIGURE 16. Two graphs with the same polynomials $M(G; x)$.

invariant for all triangle-free connected graphs on $p \leq 10$ nodes. That is, $M(G; x)$ is distinct for each triangle-free connected graph on p nodes for $p \leq 10$.

5.1. The bipartite graph $K_{2,p}$: a triangle-free, non-tree example. We now give explicit formulae for the number and sizes of the MECs on the complete bipartite graph $K_{2,p}$. For convenience, we consider the vertex set of $K_{2,p}$ to be two distinguished nodes $\{a, b\}$ together with the remaining p nodes, labeled by $[p]$, which are collectively referred to as the *spine* of $K_{2,p}$. This labeling of $K_{2,p}$ is depicted on the left in Figure 17. It is easy to see that the maximum number of immoralities is given by orienting the edges such that all edge heads are at the nodes a and b . This results in $m(K_{2,p}) = 2\binom{p}{2}$. Next, we compute a closed-form formula for the number of MECs for $K_{2,p}$.

Theorem 5.1. *The number of MECs with skeleton $K_{2,p}$ is*

$$M(K_{2,p}) = \sum_{k=0}^p \binom{p}{k} (2^{p-k} - 1 + 2^k - k) - p2^{p-1}.$$

Proof. To arrive at the desired formula, we divide the problem into three cases:

- (a) The number of immoralities at node b is $\binom{p}{2}$.
- (b) The number of immoralities at node b is strictly between 0 and $\binom{p}{2}$.
- (c) There are no immoralities at node b .

Notice that cases (a) and (b) have a natural interpretation via the indegree at node b of the essential graph of the corresponding MECs. If the indegree at b is two or more, all edges adjacent to b are essential, and the number of immoralities at node b is given by its indegree. Thus, we can rephrase cases (a) and (b) as follows:

- (a) The indegree of node b in the essential graph of the MEC is p .
- (b) The indegree of node b in the essential graph of the MEC is $1 < k < p$.

In case (a), the MEC is determined exactly by the MEC on the star with center node a and p edges. One can easily check (this was also proven as part of Theorem 2.4) that this yields $2^p - p$ MECs.

Case (b) is more subtle. First, assume that the indegree at node b is $1 < k < p$, and the arrows with head b have the tails $\{1, 2, \dots, k\} \subset [p]$. Then the remaining arrows adjacent to b are all directed outwards with heads $\{k+1, \dots, p\}$. Notice that no immoralities can happen at nodes $[k]$ along the spine, but some may occur at the nodes $[p] \setminus [k]$. If there are no such immoralities, then node a has indegree p , otherwise the essential graph would contain a directed 4-cycle. Similarly, if, without loss of generality, we denote the nodes in $[p] \setminus [k]$ that are the heads of immoralities by $\{k+1, k+2, \dots, k+s\}$ for $0 \leq s < p-k$, then the nodes $k+s+1, \dots, p$ are tails of the arrows adjacent to node a . Thus, if the number of immoralities with heads in $[p] \setminus [k]$ is $0 \leq s < p-k$, then the immoralities with heads at node a are completely determined. Therefore, each s -subset of $[p] \setminus [k]$ yields a single MEC. Figure 17 depicts an example of one such choice of immoralities. We start by selecting the arrows to form immoralities at node b which forces the remaining arrows at b to point towards the spine. We then select some of these to form immoralities at the spine, and this forces the remaining arrows to be directed inwards towards a .

However, if $s = p-k$, the star induced by nodes $\{a, 1, 2, \dots, k\}$ determines the MECs. This yields $2^k - k$ classes (see again Theorem 2.4). In total, for case (b)

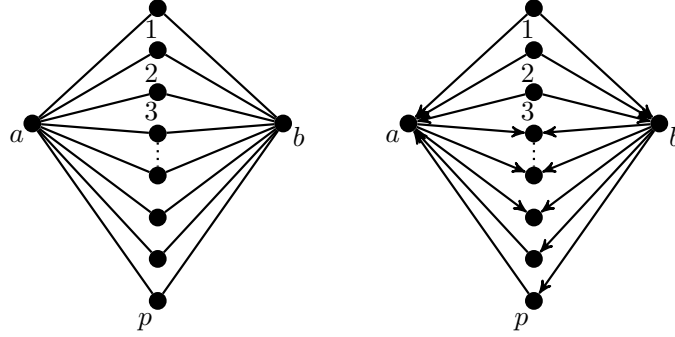


FIGURE 17. The graph $K_{2,p}$ is depicted on the left, and one of the essential graphs counted in the proof of Theorem 5.1 is depicted on the right.

the number of MECs is

$$\sum_{k=2}^{p-1} \binom{p}{k} (2^{p-k} - 1 + 2^k - k).$$

In case (c), we consider the case when there are no immoralities at node b , and we count via placement of immoralities along the spine. There are 2^p ways to place immoralities along the spine, one for each subset of $[p]$. Suppose the immoralities along the spine have the heads $\{1, 2, \dots, k\}$ for $k < p - 1$ (the cases $k = p - 1$ and $k = p$ are considered separately). Then the remaining immoralities can happen at node a . However, if there is an immorality with head at node a then all other arrows adjacent to a are essential, some of which may point towards the spine with heads in the set $[p] \setminus [k]$. Since there are no immoralities with head in the set $[p] \setminus [k]$, then any such outward pointing arrow is part of a directed path from a to b . However, since there are no immoralities at node b , there can be at most one such directed path. The presence of any such directed path forces a directed 4-cycle since $k < p - 1$. Therefore, for $k < p - 1$ the nodes $\{k + 1, \dots, p\}$ must be tails of arrows oriented towards node a , thereby yielding only a single MEC. Since $k = p$ and $k = p - 1$ also yield only a single MEC, case (c) yields a total of 2^p classes. Combing the total number of MECs counted for each of these cases yields the desired formula. $\square$

Using the case-by-case analysis from the proof of Theorem 5.1 we can count the number of MECs with skeleton $K_{2,p}$ of each possible size. Similarly, one can also recover the statistics $m_k(K_{2,p})$ from this proof. However, to avoid overwhelming the reader with formulae, we omit the expressions for $m_k(K_{2,p})$.

Corollary 5.2. *The possible sizes of a MEC with skeleton $K_{2,p}$ and the number of classes having each size is as follows:*

Class size	Number of Classes
1	$2 + \sum_{k=2}^{p-1} \binom{p}{k} 2^{p-k}$
2	$2 + \binom{p}{2}$
$3 \leq k \leq p - 1$	$1 + \binom{p}{2}$
p	2

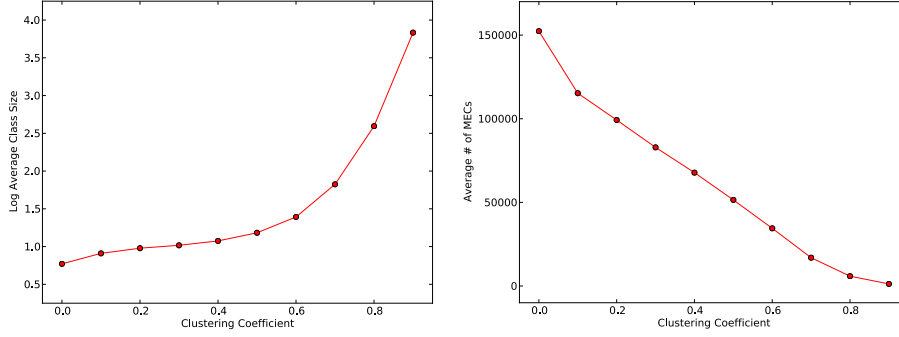


FIGURE 18. Clustering coefficient as compared to log average class size and the average number of MECs for connected graphs with $p \leq 10$ nodes and 25 edges.

Proof. Recall the case analysis from the proof of Theorem 5.1. In case (a) all MECs are size 1 except for one which is size p . This yields $2^p - p - 1$ classes of size one and one class of size p . In case (b), all MECs have size 1, unless $s = p - k$ and there are no immoralities at node a , in which case the class size is k . This yields $\binom{p}{k}$ classes of size k for $1 < k < p$, and

$$\sum_{k=2}^{p-1} \binom{p}{2} (2^{p-k} - 1)$$

classes of size 1. In case (c), all MECs have size $p - k$ for $0 \leq k < p - 1$. When $k = p - 1$, we get a single class of size 2, and when $k = p$ we get one more class of size 1. The total number of MECs of size 1 is then

$$\begin{aligned} (2^p - p - 1) + 1 + \sum_{k=2}^{p-1} \binom{p}{k} (2^{p-k} - 1) &= (2^p - p - 1) + 1 + \sum_{k=2}^{p-1} \binom{p}{k} 2^{p-k} - \sum_{k=2}^{p-1} \binom{p}{k}, \\ &= 2^p + 2 + \sum_{k=2}^{p-1} \binom{p}{k} 2^{p-k} - \sum_{k=0}^p \binom{p}{k}, \\ &= 2 + \sum_{k=2}^{p-1} \binom{p}{k} 2^{p-k}. \end{aligned}$$

The other formulae are quickly realized from the above arguments. $\square$

5.2. Skeletal structure in relation to the number and size of MECs. We now take a computational approach to analyzing the number and size of MECs on triangle-free graphs with respect to their skeletal structure. The data analyzed here was collected using the program described in [39], and this program can be found at https://github.com/aradha/mec_generation_tool. The results of [39], and those provided in the previous sections of this paper, indicate that the number and distribution of high degree nodes in a triangle-free graph dictate the size and number of MECs allowable on the skeleton. In this section, we parse these observations in terms of the data collected via our computer program.

Recall that the (*global*) *clustering coefficient* of a graph G is defined as the ratio of the number of triangles in G to the number of connected triples of vertices in G .

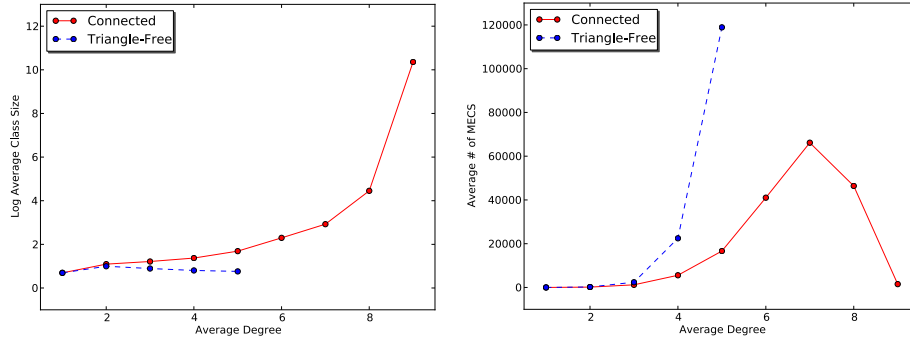


FIGURE 19. Average degree versus log average class size and average number of MECs for all graphs and triangle-free graphs on 10 nodes.

The clustering coefficient serves as a measure of how much the nodes in G cluster together. Figure 18 presents two plots: one compares the clustering coefficient to the log average class size and the other compares it to the average number of MECs. This data is taken over all connected graphs on $p \leq 10$ nodes with 25 edges (to achieve a large number of MECs). As we can see, the average class size grows as the clustering coefficient increases. This is to be expected, since an increase in the number of triangles within the DAG should correspond to an increase in the size of the chain components of the essential graph. On the other hand, the average number of MECs decreases with respect to the clustering coefficient, which is to be expected given that the class sizes are increasing. This decrease in the average number of MECs empirically captures the intuition that having many triangles in a graph results in fewer induced 3-paths, which represent the possible choices for distinct MECs with the same skeleton.

Figure 19 presents a pair of plots, the first of which compares the average degree of the underlying skeleton of the DAG to the log average class size of the associated MEC. The second plot compares the average degree of the skeleton to the average number of MECs it supports. Both plots present one curve for all connected graphs and a second curve for triangle-free graphs on 10 nodes. For connected graphs on 10 nodes the left-most plot shows a strict increase in the log average MEC class size as the average degree of the nodes in the underlying skeleton increases. This is to be expected since graphs with a higher average degree are more likely to contain larger chain components. On the other hand, the average class size for triangle-free graphs increases for average degree up until approximately 2.0, and then shows a steady decrease for larger average degree. Since the average degree of a tree on p nodes is $2 - \frac{2}{p}$, this suggests that the largest MECs amongst triangle-free graphs have skeleta being trees. As such, the bounds developed in Section 3 of this paper can be, heuristically, thought to apply more generally to all triangle-free graphs.

The right-most plot in Figure 19 describes the relationship between average degree and the average number of MECs for all connected graphs and triangle-free graphs on 10 nodes. We see from this that in the setting of all connected graphs, the skeleta with the largest average number of MECs appear to have average degree 7, whereas in the triangle-free setting, the higher the average degree the more

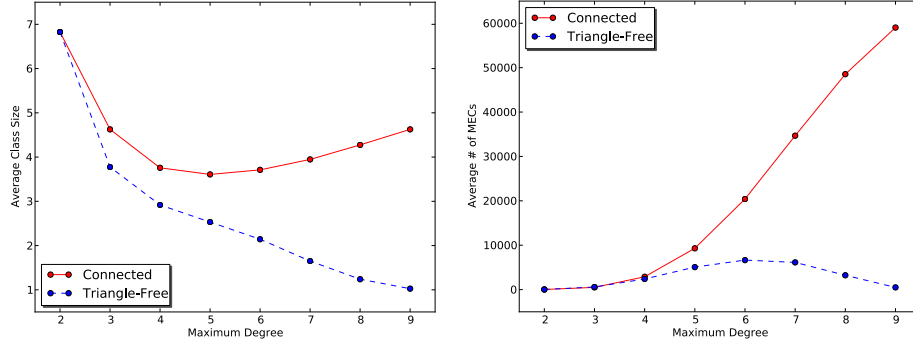


FIGURE 20. Maximum degree versus log average class size and average number of MECs for all graphs and triangle-free graphs on 10 nodes.

equivalence classes the skeleta can support. This supports the intuition that the more high degree nodes there are in a triangle-free graph, the more equivalence classes the graph can support.

The left-most plot in Figure 20 depicts the relationship between the maximum degree of a node in a skeleton and the average class size on the skeleton for all connected graphs and for triangle-free graphs on 10 nodes. For all graphs, the relationship appears to be almost linear beginning with maximum degree 5, suggesting that average class size grows linearly with the maximum degree of the underlying skeleton. This growth in class size is due to the introduction of many triangles as the maximum degree grows. On the other hand, in the triangle-free setting we actually see a decrease in average class size as the maximum degree grows, which empirically reinforces this intuition.

The right-most plot in Figure 20 records the relationship between the maximum degree of a node in a skeleton and the average number of MECs supported by that skeleton for all connected graphs and triangle-free graphs on at most 10 nodes. For all graphs, we see that the average number of MECs grows with the maximum degree of the graphs, and this growth is approximately exponential. In the triangle-free setting, the average number of MECs appears to be unimodal, but would be increasing if we considered also all graphs on $p > 10$. For triangle-free graphs there is only one graph with maximum degree 9, namely the star $G_1(9)$, where the number of MECs is $2^9 - 9$. For connected graphs the average number of MECs is pushed up by those cases consisting of a complete bipartite graph where in addition one node is connected to all other nodes.

The final plot of interest is in Figure 21, and it shows the relationship between MEC size and the ratio of the number of immoralities in the MEC to the number of induced 3-paths in the skeleton for all connected graphs and triangle-free graphs on 10 nodes. That is, it shows the relationship between the class size and how many of the potential immoralities presented by the skeleton are used by the class. It is interesting to note that, in the triangle-free setting, as the class size grows, this ratio appears to approach 0.3, suggesting that most large MECs use about a third of the possible immoralities in triangle-free graphs. In the connected graph setting, as the class size grows, we see a steady decrease in the value of this ratio. This

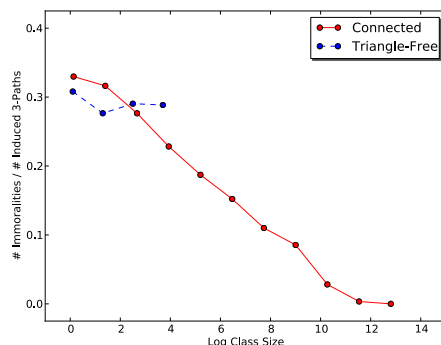


FIGURE 21. Class size versus the ratio of the number of immoralities to the number of induced 3-paths in all MECs on 10 nodes.

supports the intuition that a larger class size corresponds to an essential graph with large chain components and few immoralities.

Acknowledgements. We wish to thank Brendan McKay for some helpful advice in the use of the programs `nauty` and `Traces` [33]. Adityanarayanan Radhakrishnan was supported by ONR (N00014-17-1-2147). Liam Solus was partially supported by an NSF Mathematical Sciences Postdoctoral Research Fellowship (DMS - 1606407). Caroline Uhler was partially supported by DARPA (W911NF-16-1-0551), NSF (1651995), and ONR (N00014-17-1-2147).

REFERENCES

- [1] P. A. Aguilera, A. Fernández, R. Fernández, R. Rumi, and A. Salmerón. *Bayesian networks in environmental modelling*. Environmental Modelling & Software 26.12 (2011): 1376–1388.
- [2] Y. Alavi, P. J. Malde, A. J. Schwenk, and P. Erdős. *The vertex independence sequence of a graph is not constrained*. Congr. Numer 58 (1987): 15–23.
- [3] S. A. Andersson, D. Madigan, and M. D. Perlman. *A characterization of Markov equivalence classes for acyclic digraphs*. The Annals of Statistics 25.2 (1997): 505–541.
- [4] V. C. Barbosa and J. L. Szwarcfiter. *Generating all the acyclic orientations of an undirected graph*. Information Processing Letters 72.1 (1999): 71–74.
- [5] M. Bašić and A. Ilić. *On the clique number of integral circulant graphs*. Applied Mathematics Letters 22.9 (2009): 1406–1411.
- [6] B. Bollobás. *The independence ratio of regular graphs*. Proceedings of the American Mathematical Society (1981): 433–436.
- [7] B. Braun and L. Solus. *Shellability, Ehrhart theory, and r -stable hypersimplices*. Submitted to Journal of Combinatorial Theory Series A. ArXiv preprint arXiv:1408.4713 (2015).
- [8] J. Brown and R. Hoshino. *Independence polynomials of circulants with an application to music*. Discrete Mathematics 309.8 (2009): 2292–2304.
- [9] J. Brown and R. Hoshino. *Well-covered circulant graphs*. Discrete Mathematics 311.4 (2011): 244–251.
- [10] P. Cain. *Decomposition of complete graphs into stars*. Bulletin of the Australian Mathematical Society 10.01 (1974): 23–30.
- [11] J. M. Carraher, D. Galvin, S. G. Hartke, A. J. Radcliff, and D. Stolee. *On the independence ratio of distance graphs*. ArXiv preprint arXiv:1401.7183 (2014).
- [12] D. M. Chickering. *Learning equivalence classes of Bayesian-network structures*. Journal of Machine Learning Research 2 (2002): 445–498.
- [13] E. Cohen and M. Tarsi. *NP-completeness of graph decomposition problems*. Journal of Complexity 7.2 (1991): 200–212.

- [14] M. Drton, B. Sturmfels, and S. Sullivan. *Lectures on Algebraic Statistics*. Vol. 39. Springer Science & Business Media, 2008.
- [15] J. A. Ellis-Monaghan and C. Merino. *Graph polynomials and their applications I: The Tutte polynomial*. Structural Analysis of Complex Networks. Birkhäuser Boston, 2011. 219–255.
- [16] J. A. Ellis-Monaghan and C. Merino. *Graph polynomials and their applications II: Interrelations and interpretations*. Structural Analysis of Complex Networks. Birkhäuser Boston, 2011. 257–292.
- [17] N. Friedman, M. Linial, I. Nachman and D. Peter. *Using Bayesian networks to analyze expression data*. Journal of Computational Biology 7 (2000): 601–620.
- [18] M. R. Garey and D. S. Johnson. *Computers and intractability: a guide to the theory of NP-completeness*. A Series of Books in the Mathematical Sciences. WH Freeman and Company, New York, NY 25.27 (1979): 141.
- [19] S. B. Gillispie. *Formulas for counting acyclic digraph Markov equivalence classes*. Journal of Statistical Planning and Inference 136.4 (2006): 1410–1432.
- [20] S. B. Gillispie and M. D. Perlman. *Enumerating Markov equivalence classes of acyclic digraph models*. Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., 2001.
- [21] N. Hamada, H. Ikeda, S. Shiga-eda, K. Ushio, and S. Yamamoto. *On claw-decomposition of complete graphs and complete bigraphs*. Hiroshima Mathematical Journal 5.1 (1975): 33–42.
- [22] Y. He, J. Jia, and B. Yu. *Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs*. J. Mach. Learn. Res 16 (2015): 2589–2609.
- [23] Y. He and B. Yu. *Formulas for counting the sizes of Markov equivalence classes of directed acyclic graphs*. ArXiv preprint arXiv: <https://arxiv.org/pdf/1610.07921.pdf> (2016).
- [24] R. Hoshino. *Independence polynomials of circulant graphs*. Library and Archives Canada, 2008.
- [25] R. M. Karp *Reducibility among combinatorial problems*. Complexity of Computer Computations. Springer US (1972): 85–103.
- [26] T. Koshy. *Fibonacci and Lucas numbers with applications*. Vol. 51. John Wiley & Sons, 2011.
- [27] V. E. Levit and E. Mandrescu. *On well-covered trees with unimodal independence polynomials*. Congressus Numerantium (2002): 193–202.
- [28] V. E. Levit and E. Mandrescu. *On unimodality of independence polynomials of some well-covered trees*. Discrete Mathematics and Theoretical Computer Science. Springer Berlin Heidelberg, 2003. 237–256.
- [29] V. E. Levit and E. Mandrescu. *The independence polynomial of a graph – a survey*. Proceedings of the 1st International Conference on Algebraic Informatics. Vol. 233254. 2005.
- [30] C. Lin, and T-W. Shyu. *A necessary and sufficient condition for the star decomposition of complete graphs*. Journal of Graph Theory 23.4 (1996): 361–364.
- [31] J. H. van Lint and R. M. Wilson. *A Course in Combinatorics*. Cambridge University Press, 2001.
- [32] J. L. Martin, M. Morin, and J. D. Wagner. *On distinguishing trees by their chromatic symmetric functions*. Journal of Combinatorial Theory, Series A 115.2 (2008): 237–253.
- [33] B. D. McKay and A. Piperno. *Practical graph isomorphism, II*. Journal of Symbolic Computation 60 (2014): 94–112.
- [34] C. Meek. *Causal inference and causal explanation with background knowledge*. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (1995): 403–410.
- [35] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.
- [36] S. Poljak. *A note on stable sets and colorings of graphs*. Commentationes Mathematicae Universitatis Carolinae 15.2 (1974): 307–309.
- [37] H. Prodinger and R. F. Tichy. *Fibonacci numbers of graphs*. Fibonacci Quarterly 20.1 (1982): 16–21.
- [38] H. Prodinger and R. F. Tichy. *Fibonacci numbers of graphs. II*. Fibonacci Quarterly 21.3 (1983): 219–229.
- [39] A Radhakrishnan, L. Solus, and C. Uhler. *Counting Markov equivalence classes by number of immoralities*. To appear in the Proceedings of the 2017 Conference on Uncertainty in Artificial Intelligence (2017).
- [40] J. M. Robins, M. A. Hernán and B. Brumback. *Marginal structural models and causal inference in epidemiology*. Epidemiology 11.5 (2000): 550–560.

- [41] N. J. Sloane. *The On-Line Encyclopedia of Integer Sequences*. (2003).
- [42] L. Solus, Y. Wang, C. Uhler, and L. Matejovicova. *Consistency guarantees for permutation-based causal inference algorithms*. Preprint available at: <https://arxiv.org/abs/1702.03530> (2017).
- [43] P. Spirtes, C. N. Glymour and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, 2001.
- [44] R. P. Stanley. *A symmetric function generalization of the chromatic polynomial of a graph*. Advances in Mathematics 111.1 (1995): 166-194.
- [45] B. Steinsky. *Enumeration of labelled chain graphs and labelled essential directed acyclic graphs*. Discrete Mathematics 270.1 (2003): 267-278.
- [46] M. Tarsi. *Decomposition of complete multigraphs into stars*. Discrete Mathematics 26.3 (1979): 273-278.
- [47] K. Ushio. *G-designs and related designs*. Discrete Mathematics 116.1 (1993): 299-311.
- [48] K. Ushio, S. Tazawa and S. Yamamoto. *On claw-decomposition of a complete multipartite graph*. Hiroshima Mathematical Journal 8.1 (1978): 207-210.
- [49] T. Verma and J. Pearl. *An algorithm for deciding if a set of observed independencies has a causal explanation*. Proceedings of the Eighth International Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., 1992.
- [50] S. Wagner. *Asymptotic enumeration of extensional acyclic digraphs*. Algorithmica 66.4 (2013): 829-847.

LABORATORY FOR INFORMATION AND DECISION SYSTEMS,, AND INSTITUTE FOR DATA, SYSTEMS,
AND SOCIETY, MIT, CAMBRIDGE, MA, USA
E-mail address: `aradha@mit.edu`

KTH ROYAL INSTITUTE OF TECHNOLOGY, STOCKHOLM, SWEDEN
E-mail address: `solus@kth.se`

LABORATORY FOR INFORMATION AND DECISION SYSTEMS,, AND INSTITUTE FOR DATA, SYSTEMS,
AND SOCIETY, MIT, CAMBRIDGE, MA, USA
E-mail address: `cuhler@mit.edu`