Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

EQUIVALENCE CLASSES OF STAGED TREES

By Christiane Görgen and Jim Q. Smith

University of Warwick

In this paper we give a complete characterization of the equivalence classes of CEGs or equivalently of staged trees. This model class cannot be unambiguously indexed by its graphical properties. However, we are able to show that a polynomial defined on an underlying graph codes all relevant characteristics and is common to all representations of the same model. Furthermore, simple transformations on that polynomial enable us to traverse the statistical equivalence class of these graphs. So one can design efficient algorithms over the classes. We illustrate our results throughout the paper, finishing with a real analysis of the implicit dependence relationships found in [6].

1. Introduction.

The *Chain Event Graph (CEG)* is a discrete statistical model based on a graphical description given by an event tree [30]. CEGs have now successfully led statistical inference in a whole range of domains [2, 6, 12, 33, 34, 35]. However, a formal analysis of the statistical properties of this class of models is long overdue.

In this paper, it will be most convenient to represent a CEG model by a corresponding *staged tree* [30]. From this coloured graph we can read a parametrization rule given by the multiplication of transition probabilities along root-to-leaf paths. Two staged trees are said to be *statistically equivalent* if their parametrization rules parametrize the same model, see Section 2.

The study of these statistical equivalence classes is an important one. The first reason for this is computational: CEGs constitute a massive model space to explore. By identifying a single representative within an equivalence class of model representations and a priori selecting across these representatives rather than the full class, we can dramatically reduce the search effort across this space. The second reason concerns coherence: when adopting a Bayesian approach in model selection, [16] and others have argued that two statistically equivalent models (i.e. those always giving the same likelihood) should be given the same prior distribution over its parameters. To apply this principle, it is essential to know when two CEGs make the same distributional assertions. The third reason is inferential: Just like a Bayesian

MSC 2010 subject classifications: Primary 60E05, 60K35; secondary 62E99

Keywords and phrases: Staged Trees, Chain Event Graphs, Probability Trees, Algebraic Statistics

network (BN), a CEG or staged tree has a natural causal extension [6, 32]. So, in particular, causal discovery algorithms can be applied to CEGs to elicit a putative causal ordering between various associated variables. Clearly a necessary condition for a causal deduction to be made is that this deduction is invariant to the choice of one representative within a statistical equivalence class. So again we need to be able to identify equivalence classes of a hypothesized causal CEG in order to perform these algorithms.

Now, unlike for BNs, where model representations making equivalent distributional assumptions can be elegantly characterized through their sharing the same *essential graph* [1, 16], sadly no such common representation is available for staged trees or CEGs. However, we show here that we can instead specify staged trees as representations of a set of monomials together with linear constraints which characterize the model. This then provides a natural algebraic index for a class of equivalent staged trees to be used as an analogue of the essential graph classifying equivalence classes of BNs. Because staged tree models include discrete BN models as a special case, our characterization also gives an alternative to the ansatz adopted by [13].

Our central theorem, presented in Section 3, is based on two main findings: First, the *interpolating polynomial* of a staged tree can capture certain context specific independence structures which are invariant to a class of graphical transformations we call *swaps*. These transformations are analogous to arc reversals sometimes applied to BN models [27]. Second, by substituting various monomial terms of the interpolating polynomial into single factors we can often simplify our representation to capture only its substantive structure. Within our development this corresponds to what we call here a *resize* operator on the staged tree. We show later that in the context of decomposable BNs, this operation is analogous for example to the transformation of a directed acyclic graph (DAG) into a junction tree [17]. Swaps and resizes enable us to meaningfully incrementally traverse the class of statistically equivalent staged tree representations of a given model. We are able to show that between every two statistically equivalent staged trees there is a map which is a composition of these operators. Statistical equivalence classes of staged tree and CEG models are thus fully characterized through simple relationships between their interpolating polynomials.

In Section 4, a full characterization of the statistical equivalence class and a putative causal interpretation of the staged tree representing the Christchurch Health and Development Study [11] rounds off the analysis. We end the paper with a brief discussion about how this work is currently being used and extended.

2. Statistical equivalence for staged trees. In this paper we study properties of parametric statistical models based on a graphical representation given by a probability tree [3, 26, 28, 30]. We will treat the probability tree not only as some easily interpretable picture but also as a directed graphical model in its own right. To properly study equivalence classes of these models, we first need to tighten the formalism introduced in [30].

2.1. *Discrete parametric models and probability trees.* We begin by recalling a technical definition of a discrete parametric model. This enables us to discuss those discrete models which can be represented by probability trees.

Let Ω always denote a finite space with $n \geq 2$ *atoms* $\omega \in \Omega$. Write $p_\theta : \Omega \to (0,1)$ to denote a strictly positive probability mass function which depends on a vector of parameters $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$. Denote the vector of values of that function by $\boldsymbol{p}_\theta = \big(p_\theta(\omega) \mid \omega \in \Omega\big)$ and call each of its components $p_\theta(\omega)$, $\omega \in \Omega$, an *atomic probability.* We can then define a discrete *parametric statistical model* on Ω as a subset of the open $n-1$ dimensional probability simplex

$$\mathbb{P}_\Psi = \{\boldsymbol{p}_\theta \mid \theta \in \Theta\} \subseteq \Delta^\circ_{n-1} \tag{2.1}$$

where $\Delta^\circ_{n-1} = \big\{p \in \mathbb{R}^n \mid \sum_{i\in[n]} p_i = 1,\ p_i \in (0,1) \text{ for all } i \in [n]\big\}$ and where $[n] = \{1, \ldots, n\}$ [9]. The index Ψ in $\mathbb{P}_\Psi$ is a bijective map

$$\Psi :\ \Theta \to \mathbb{P}_\Psi, \quad \theta \mapsto \boldsymbol{p}_\theta \tag{2.2}$$

called a *parametrization* of the model $\mathbb{P}_\Psi$.

We say that Ψ is a *monomial parametrization* if every component of its image is a monomial $\Psi_i(\theta) = \theta_1^{\alpha_{i,1}} \cdots \theta_d^{\alpha_{i,d}}$ with exponent $\alpha_i \in \mathbb{N}_0^d$, $i \in [n]$, where $\theta = (\theta_1, \ldots, \theta_d)$. A monomial parametrization is called *multilinear* if $\alpha_i \in \{0,1\}^d$ for all $i \in [n]$. We henceforth call two parametrizations Ψ and Φ of the same model, so $\mathrm{im}(\Psi) = \mathbb{P}_\Psi = \mathbb{P}_\Phi = \mathrm{im}(\Phi)$, *statistically equivalent.*

The following terminology will enable us to discuss parametric statistical models whose parametrization can be read from a particular graph. The characterization of statistical equivalence classes of these graphical models and its implications will then be the focus of this paper.

A finite graph $\mathcal{T} = (V, E)$ with vertex set V and edge set $E \subseteq V \times V$ is called a *tree* if it is connected and has no cycles [28]. In a *directed* tree, each $e = (v, v') \in E$ is a pair of ordered vertices. We call vertices $\mathrm{pa}(v) = \{v' \mid \exists (v', v) \in E\}$ the *parents* of $v \in V$ and $\mathrm{ch}(v) = \{v' \in V \mid \exists (v, v') \in E\}$ the set of *children* of $v \in V$. A vertex $v_0 \in V$ without parents is called a *root* of the tree and vertices without children are called *leaves*. We use the term *root-to-leaf path* for a directed sequence of edges $\lambda = (e \mid e \in E(\lambda))$ with $E(\lambda) \subseteq E$, where one edge $e = (v, v')$ precedes another $e' = (w, w')$ only if $v' = w$. A *subpath* within a directed tree is then a connected subsequence of a root-to-leaf path.

Recall that within a tree model (as defined below), every root-to-leaf path represents an atom in a given sample space and depicts one possible history of a unit in a population passing through the tree. Every vertex $v \in V$ denotes a situation that such a unit might find itself in during that progress, and every edge $e = (v, v') \in E$ denotes the possibility of passing from one situation v to the next v'.

We call a directed tree an *event tree* if all vertices except for one unique root have exactly one parent and each parent which is not a leaf has at least two children. Then for any unit in the population there are always at least two possible unfoldings from every situation it might pass through.

We denote the set of all root-to-leaf paths of an event tree by $\Lambda(\mathcal{T})$. The power set of the set of root-to-leaf paths is called the *path sigma-algebra* of the tree, denoted $\sigma(\mathcal{T}) = \mathcal{P}(\Lambda(\mathcal{T}))$. For a fixed $v \in V$ or $e \in E$ we define *vertex-* or *edge-centered events* within the path sigma-algebra as

$$\Lambda(v) = \big\{\lambda \in \Lambda(\mathcal{T}) \mid \text{there is } (\cdot, v) \in E(\lambda)\big\}, \tag{2.3a}$$

$$\Lambda(e) = \big\{\lambda \in \Lambda(\mathcal{T}) \mid \text{there is } e \in E(\lambda)\big\}, \tag{2.3b}$$

respectively, and set $\Lambda(v_0) = \Lambda(\mathcal{T})$. In tree models, the set of all root-to-leaf paths going through one fixed vertex (or edge) is the set of all atoms for which that situation happens. These sets are called *Moivrean events* in [28].

We call a pair $\mathcal{F}_v = (v, E(v))$ where $E(v) = \{(v, v') \in E \mid v' \in \mathrm{ch}(v)\}$ and $v \in V$ a *floret*. If v is a leaf then $E(v) = \emptyset$ and $\mathcal{F}_v$ is an empty floret. The notion of root-to-leaf paths in an event tree allows us to order florets, edges $e, e' \in E$ and vertices $v, v' \in V$ (and hence events as in (2.3a) and (2.3b)) in the following way: We say that $e \prec e'$, $v \prec v'$ or $\mathcal{F}_v \prec \mathcal{F}_{v'}$ if and only if every root-to-leaf path $\lambda \in \Lambda(e) \cap \Lambda(e')$ or $\lambda \in \Lambda(v) \cap \Lambda(v')$ is a sequence of edges containing e before e' or $(v, \cdot)$ before $(v', \cdot)$, respectively. Thus, from [30], $\prec$ defines a pre-order on an event tree. Tree models therefore admit a natural directionality and are particularly useful if a model class needs to express a potential ordering of events, rather than of random variables.

Definition 1 (Probability tree). *Let $\mathcal{T} = (V, E)$ be an event tree with a finite number of root-to-leaf paths and associate parameters $\theta(e) = \theta(v, v')$ to all edges $e = (v, v') \in E$. We call $\theta_v = \big(\theta(e) \mid e \in E(v)\big)$ a* vector of floret parameters*.*

The pair $(\mathcal{T}, \Theta_{\mathcal{T}})$ *is called a* probability tree *if* $\Theta_{\mathcal{T}} = \{\theta_v \mid v \in V\}$ *is a set of floret parameter vectors where* $\theta_v \in \Delta^{\circ}_{\#E(v)-1}$ *for all* $v \in V$*. In probability trees, we call each* $\theta(e)$*,* $e \in E$*, a* primitive probability*.*

Primitive probabilities can be thought of as (conditional) transition probabilities along root-to-leaf paths, as we outline below. Importantly, the constraint $\theta_v \in \Delta^{\circ}_{\#E(v)-1}$ for all $v \in V$ ensures that primitive probabilities are positive and those belonging to the same floret sum to unity. Henceforth, we assume these probability simplices are open in order to avoid various distracting technical issues concerning boundary cases.

Let $(\mathcal{T}, \Theta_{\mathcal{T}})$ be a given probability tree, $\mathcal{T} = (V, E)$. Denote the product of all primitive probabilities along a root-to-leaf path $\lambda \in \Lambda(\mathcal{T})$ by

$$\pi_{\theta,\mathcal{T}}(\lambda) = \prod_{e \in E(\lambda)} \theta(e) \tag{2.4}$$

where $\theta = \big(\theta(e) \mid e \in E\big)$ is a vector of all primitive probabilities. We show below that $\pi_{\theta,\mathcal{T}}$ is a probability mass function on a probability space $(\Lambda(\mathcal{T}), \sigma(\mathcal{T}), \pi_{\theta,\mathcal{T}})$ associated to $(\mathcal{T}, \Theta_{\mathcal{T}})$ and that (2.4) induces a monomial parametrization as in (2.2).

By Definition 1, $\pi_{\theta,\mathcal{T}}(\lambda) \in (0,1)$ for every $\lambda \in \Lambda(\mathcal{T})$. Moreover, the probabilities of all root-to-leaf paths sum to unity:

Lemma 1. *Let* $(\mathcal{T}, \Theta_{\mathcal{T}})$ *be a probability tree and* $\mathcal{T} = (V, E)$*. Then the constraint that* $\theta_v \in \Delta^{\circ}_{\#E(v)-1}$ *for all* $v \in V$ *is true if and only if*

$$\sum_{\lambda \in \Lambda(\mathcal{T})} \pi_{\theta,\mathcal{T}}(\lambda) = 1. \tag{2.5}$$

This result is immediately apparent from substituting subsums in the polynomial in (2.5) which sum to unity by definition.

The importance of Lemma 1 will become apparent in Sections 2.2 and 3.1 where we search for different probability trees that have the same probability mass function $\pi_{\theta,\mathcal{T}}$. In particular, the local floret structure on an underlying graph can vary provided that the root-to-leaf path condition is satisfied.

From the above, the map $\pi_{\theta,\mathcal{T}} : \sigma(\mathcal{T}) \to [0,1]$, $A \mapsto \sum_{\lambda \in A} \pi_{\theta,\mathcal{T}}(\lambda)$ is a strictly positive probability measure. Thus, $\boldsymbol{\pi}_{\theta,\mathcal{T}} = \big(\pi_{\theta,\mathcal{T}}(\lambda) \mid \lambda \in \Lambda(\mathcal{T})\big)$ is a vector of atomic probabilities and

$$\mathbb{P}_{(\mathcal{T},\Theta_{\mathcal{T}})} = \Big\{ \boldsymbol{\pi}_{\theta,\mathcal{T}} \mid \theta \in \bigtimes_{v \in V} \Delta^{\circ}_{\#E(v)-1} \Big\} \subseteq \Delta^{\circ}_{\#\Lambda(\mathcal{T})-1} \tag{2.6}$$

denotes a parametric model. We call (2.6) a *(probability) tree model* and say that the elements in $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}$ *factorize according to* $\mathcal{T}$, analogous to [18] where in a BN model distributions factor according to a DAG. Note that a tree model is always indexed by one possible representation $(\mathcal{T},\Theta_\mathcal{T})$.

Tree models are parametric models with a parametrization as in (2.2),

$$\begin{aligned}\Psi_\mathcal{T}:\quad & \bigtimes_{v\in V}\Delta^\circ_{\#E(v)-1}\to\Delta^\circ_{\#\Lambda(\mathcal{T})-1}\\ (2.7)\qquad & \big(\theta(e)\mid e\in E\big)\ \mapsto\ \Big(\prod_{e\in E(\lambda)}\theta(e)\mid\lambda\in\Lambda(\mathcal{T})\Big)\end{aligned}$$

based on (2.4). Hence, a probability tree $(\mathcal{T},\Theta_\mathcal{T})$ is a graphical representation of a monomial parametrization $\Psi_\mathcal{T}$ of a family of probability mass functions $\mathbb{P}_{\Psi_\mathcal{T}}=\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}$ which factorize according to the graph $\mathcal{T}$.

Two parametrizations $\Psi_\mathcal{S}$ and $\Psi_\mathcal{T}$ giving rise to the same tree model $\mathrm{im}(\Psi_\mathcal{S})=\mathbb{P}_{(\mathcal{S},\Theta_\mathcal{S})}=\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}=\mathrm{im}(\Psi_\mathcal{T})$ are by definition statistically equivalent. Henceforth, we will also call the two probability tree representations $(\mathcal{T},\Theta_\mathcal{T})$ and $(\mathcal{S},\Theta_\mathcal{S})$ of that model statistically equivalent. We let the symbol $[\mathcal{T},\Theta_\mathcal{T}]$ denote the set of all probability tree representations of $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}$.

We can always identify the set of root-to-leaf paths $\Lambda(\mathcal{T})$ of a probability tree $(\mathcal{T},\Theta_\mathcal{T})$ with a finite space Ω via a bijection

$$(2.8)\qquad \iota_\mathcal{T}:\ \Omega\to\Lambda(\mathcal{T}),\quad \omega\mapsto\big(e\mid e\in E(\iota_\mathcal{T}(\omega))\big)$$

where $\mathcal{T}=(V,E)$. Importantly, $\pi_{\theta,\mathcal{T}}$ then induces a measure $P_\theta=\pi_{\theta,\mathcal{T}}\circ\iota_\mathcal{T}$ on Ω which does not depend on the graph $\mathcal{T}$. We will usually use the symbol $P_\theta(\omega)$ to refer to a value in $(0,1)$ and $\pi_{\theta,\mathcal{T}}(\iota_\mathcal{T}(\omega))$ to refer to a symbolic product of parameters, $\omega\in\Omega$: see Section 3.

By the above, a tree model $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}$ thus has an underlying probability space $(\Omega,\sigma(\Omega),P_\theta)$, where $\sigma(\Omega)$ denotes a sigma-algebra on Ω. Then two probability trees $(\mathcal{T},\Theta_\mathcal{T})$ and $(\mathcal{S},\Theta_\mathcal{S})$ are statistically equivalent if and only if they induce the same underlying probability space. So in particular,

$$(2.9)\qquad \pi_{\theta,\mathcal{T}}(\iota_\mathcal{T}(\omega))=\pi_{\theta',\mathcal{S}}(\iota_\mathcal{S}(\omega))\quad\text{for all }\omega\in\Omega$$

where $\theta=\big(\theta(e)\mid e\in E\big)$ and $\theta'=\big(\theta(e')\mid e'\in E'\big)$ are the vectors of primitive probabilities for $\mathcal{T}=(V,E)$ and $\mathcal{S}=(V',E')$, respectively.

Example 1. *Let $(\mathcal{T},\Theta_\mathcal{T})$ be a probability tree with $n\in\mathbb{N}$ root-to-leaf paths and no additional constraints on the probability mass function $\pi_{\theta,\mathcal{T}}$. Then the vector $\boldsymbol{\pi}_{\theta,\mathcal{T}}\in\Delta^\circ_{n-1}$ can take any value within the probability simplex. Hence, $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}=\Delta^\circ_{n-1}$. We will call this a* saturated *tree model.*

Let $\mathcal{F} = (v_0, \{e_1, \ldots, e_n\})$ be a floret with an associated parameter vector $\theta_{\mathcal{F}} = \big(\theta(e_i) \mid i \in [n]\big) \in \Delta^{\circ}_{n-1}$. Then $(\mathcal{F}, \{\theta_{\mathcal{F}}\})$ is statistically equivalent to the saturated tree $(\mathcal{T}, \Theta_{\mathcal{T}})$ if the probabilities associated with the same atoms are identified: so $\theta(e_i) = \pi_{\theta,\mathcal{T}}(\iota_{\mathcal{T}}(\omega_i))$ for all $\iota_{\mathcal{F}}^{-1}(e_i) = \omega_i \in \Omega$, $i \in [n]$.

In Section 3.3 we show that the map $(\mathcal{T}, \Theta_{\mathcal{T}}) \mapsto \mathcal{F}$ corresponds to a substitution of the monomials $\pi_{\theta,\mathcal{T}}(\lambda_i) = \prod_{e \in E(\lambda_i)} \theta(e) \mapsto \theta(e_i)$ for $i \in [n]$, and that a floret is a graphically minimal representation of a saturated model.

A tree model does not need a priori an underlying set of random variables. However, sometimes a problem is naturally defined through the relationships between a set of prespecified random variables. When this is so, the following semantics enable us to exploit the extra information coded in these. Consider a parametric model in the *positive discrete distribution framework* [31] where a discrete probability space $(\Omega, \sigma(\Omega), P)$ has a strictly positive measure $P = P_\theta$, with $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$. Let $X = (X_1, \ldots, X_m) : \Omega \to \mathbb{X}$ be a vector of discrete random variables on that space, where $\mathbb{X} = \mathbb{X}_1 \times \ldots \times \mathbb{X}_m$ denotes a product space, $m \in \mathbb{N}$. Suppose this probability measure admits a monomial parametrization implied by

$$P_\theta(X = x) = p_\theta(x) = \prod_{i \in [k]} \theta(x_{A_i}) \quad \text{for all } x \in \mathbb{X} \tag{2.10}$$

where x_{A_i} denotes the vector $(x_j \mid j \in A_i) \in \mathbb{X}_{A_i} = \bigtimes_{j \in A_i} \mathbb{X}_j$ for index sets $A_i \subseteq [m]$, $i \in [k]$ and $k \in \mathbb{N}$ [18]. Then $\Psi : \theta \mapsto \boldsymbol{p}_\theta = \big(p_\theta(x) \mid x \in \mathbb{X}\big)$ defines a discrete parametric statistical model $\mathbb{P}_\Psi$ as in (2.1). If $\mathbb{P}_\Psi$ is also a tree model, then any of its representations $(\mathcal{T}, \Theta_{\mathcal{T}})$ can be linked to the underlying probability space via an embedding from the state space of X, rather than Ω, into the set of paths

$$\begin{aligned} \iota_{\mathcal{T}} = \iota_{\mathcal{T},A} : \quad & \mathbb{X}_1 \times \mathbb{X}_2 \times \ldots \times \mathbb{X}_m \to \Lambda(\mathcal{T}) \\ & (x_1, \ldots, x_m) \mapsto \big(e(x_{A_1}), \ldots, e(x_{A_k})\big) \end{aligned} \tag{2.11}$$

such that $\pi_{\theta,\mathcal{T}}(\iota_{\mathcal{T}}(x)) = P_\theta(X = x)$ for all $x \in \mathbb{X}$ [30].

We assume in (2.11) that those index sets which are non-empty $A_i \neq \emptyset$ are pairwise different, $A_i \neq A_j$ for $i \neq j$. However, we do not demand that their union necessarily covers all indices $1, \ldots, m$. This is because some combinations of $x_1, \ldots, x_m$ might not make sense in a modelling context. In a tree model we can use shorter root-to-leaf paths to visualize these instead of retaining redundant information in the product space $\mathbb{X}_1 \times \mathbb{X}_2 \times \ldots \times \mathbb{X}_m$ [15]. The latter approach is often necessary when encoding a problem as a BN.

If in (2.11) the set $A = \{A_1, \ldots, A_k\}$ does not depend on $x \in \mathbb{X}$ we call the staged tree X-*compatible* [6]. Then an edge's $e(x_{A_i}) = (v, v')$ meaning of "passing from situation v to v'" can be equivalently read as "$x_{A_i \setminus A^{i-1}}$ happened'" given that "$x_{A^{i-1}}$ happened before", for $A^{i-1} = \bigcup_{j \in [i-1]} A_j$ and $i \geq 2$. Thus, the primitive probabilities $\theta(x_{A_i}) = \theta(e(x_{A_i}))$ from (2.10) are *potentials* [20, 21] with a conditional or marginal meaning that depends on the graph $\mathcal{T}$ and its sum-to-1 conditions in $\Theta_{\mathcal{T}}$. Floret parameter vectors of the type $\theta_v = \big(\theta(x_{A_i}) \mid x_{A_i \setminus A^{i-1}} \in \mathbb{X}_{A_i \setminus A^{i-1}}\big)$ are then rows of conditional probability tables. Examples 3 and 5 illustrate this.

Note that the pre-order on vertices, edges and florets in an X-compatible staged tree can be translated into a total order on the components of X. See Section 4 for an application.

2.2. *Staged tree models.* Probability trees are most interesting when two or more vectors of floret parameters are hypothesized to take the same values, and the distributions $\pi_{\theta,\mathcal{T}}$ factorize according to a "coloured" graph $\mathcal{T}$. We will analyze this type of model in the remainder of this text.

Definition 2 (Staged tree). *Let $(\mathcal{T}, \Theta_{\mathcal{T}})$, $\mathcal{T} = (V, E)$, be a probability tree. We define an equivalence relation $\sim$ on $V \times V$ which relates two vertices $v \sim w$ if and only if their parameter vectors coincide $\theta_v = \theta_w$, for $v, w \in V$. Then v and w are said to be in the same* stage. *If this staging is non-trivial, $(\mathcal{T}, \Theta_{\mathcal{T}})$ is said to be a* staged tree, *otherwise a* saturated tree.

If $\Lambda(v) \cap \Lambda(w) = \emptyset$ for any related $v \sim w$, we will call $(\mathcal{T}, \Theta_{\mathcal{T}})$ square-free.

When having a preassigned set of random variables, setting floret vectors of conditional probabilities equal to each other can be interpreted as specifying a set of *context-specific* conditional independences as in [3]. Models with these types of constraints are now widely used in BN modelling, especially when the domain of application is large. In tree models, given two vertices are in the same stage and a unit arrives at one of them, the transition probabilities to all children of that vertex will not depend on which of the two vertices the unit is actually in, and will thus not depend on the way that unit took to arrive in that situation. The edge (or transition) probabilities in these stages are thus in a sense independent of their history or location in the graph [35]. More intuition on stages can be found in [30].

In Section 3, we interpret the stage structure of a probability tree as a set of linear constraints on the primitive probabilities.

Example 2. *The staged tree $(\mathcal{T}, \Theta_{\mathcal{T}})$ depicted in Fig. 1 is a simplified detail of the graph analyzed in [2]. Here, every atom is represented by a root-to-leaf path with two edges and, via an embedding $\iota_{\mathcal{T}}$, corresponds to a possible history of a child in the study [11]. The first edge of each such path*

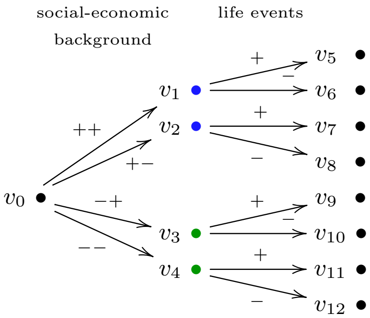

FIGURE 1. *A staged tree $(\mathcal{T}, \Theta_\mathcal{T})$, simplified version taken from [2]. We label the edges by $+$ and $-$, corresponding to 'high' and 'low', respectively. See Example 2 for a discussion.*

depicts the socio-economic background of a child, the second corresponds to a number of life events. For instance, $\lambda = ((v_0, v_2), (v_2, v_8)) \in \Lambda(\mathcal{T})$ is in one-to-one correspondence with 'high social status, low economic background, low number of life events'.

Using stages, we can embed information of the type 'if we know the social status of a child, then its number of life events does not depend on its economic situation'. In Fig. 1, the vertices $v_1 \sim v_2$ and $v_3 \sim v_4$ are then related. So the primitive probabilities of the edges of the corresponding florets are identified, $\theta(v_i, v_j) = \theta(v_{i+1}, v_{j+2})$ for $j = 2i + 3, 2i + 4$ and $i = 1, 3$.

Denote by $U_\mathcal{T}$ the by the relation $\sim$ induced partition on the vertex set, the *stage set* of $(\mathcal{T}, \Theta_\mathcal{T})$. We call elements in $\tilde{U}_\mathcal{T} = \{u \in U_\mathcal{T} \mid \#u > 1\}$ *non-trivial stages*. If $\tilde{U}_\mathcal{T} \neq \emptyset$ then $(\mathcal{T}, \Theta_\mathcal{T})$ is also called a *coloured* tree, and we actually assign all vertices in the same stage the same colour [2, 30]. We write θ_u to denote a representative of the set $\{\theta_v \mid v \in u\}$, $u \in U_\mathcal{T}$.

Every staged tree $(\mathcal{T}, \Theta_\mathcal{T})$ has a corresponding saturated tree $(\mathcal{T}, \Theta_\mathcal{T})_{\text{sat}} = (\mathcal{T}_{\text{sat}}, \Theta_{\mathcal{T}_{\text{sat}}})$ with $\mathcal{T}_{\text{sat}} = \mathcal{T}$ and $\Theta_\mathcal{T} \subseteq \Theta_{\mathcal{T}_{\text{sat}}}$ where all stages are trivial. Because stage structure imposes contraints on the underlying probability mass function, a staged tree $(\mathcal{T}, \Theta_\mathcal{T})$ represents a *submodel* $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})} \subseteq \mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})_{\text{sat}}}$ of a model represented by $(\mathcal{T}, \Theta_\mathcal{T})_{\text{sat}}$. Clearly, $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})_{\text{sat}}} = \Delta^\circ_{n-1}$, as we have seen in Example 1.

Note that the map $\theta_v \mapsto \theta_u$ for $v \in u$, $u \in U_\mathcal{T}$, induces a projection

$$\begin{aligned} \Pi : \bigtimes_{v \in V} \Delta^\circ_{\#E(v)-1} &\to \bigtimes_{u \in U_\mathcal{T}} \Delta^\circ_{\#E(u)-1} \\ \big(\theta_v \mid v \in V\big) &\mapsto \big(\theta_u \mid u \in U_\mathcal{T}\big) \end{aligned} \tag{2.12}$$

onto a usually lower dimensional space, because setting primitive probabilities equal to one another reduces the number of free parameters. $E(u)$ in (2.12) denotes the edge set $E(v)$ of one fixed representative $v \in u$. For instance, in Example 2 Π maps a vector of all primitive probabilities from $\Delta^\circ_{4-1} \bigtimes_{i\in[4]} \Delta^\circ_{2-1}$ onto the space $\Delta^\circ_{4-1} \times \Delta^\circ_{2-1} \times \Delta^\circ_{2-1}$.

The staging of a probability tree can respect certain symmetries: A vertex $v \in V$, in $\mathcal{T} = (V, E)$, is said to be at *level* i of the tree $\mathcal{T}$ if the subpath from v_0 to v has i edges, $i \in \mathbb{N}$. We call a staged tree *stratified* if all vertices in the same stage are also at the same level. The symbol $[\mathcal{T}, \Theta_\mathcal{T}]_{\text{strata}} \subseteq [\mathcal{T}, \Theta_\mathcal{T}]$ denotes the set of statistically equivalent stratified tree representations of a model $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}$. This class is amenable to various fast search algorithms [4, 6].

In particular, an X-compatible staged tree is stratified only if its stage constraints are of the form

$$\theta(x_A) = \theta(x'_A) \quad \text{for some } x_A, x'_A \in \mathbb{X}_A \tag{2.13}$$

where $A \subseteq [m]$. This is the case for instance in (context-specific) BN models. Importantly, the stratification constraint (2.13) can be used to prevent an identification of primitive probabilities which might not make sense in a modelling context: see Section 4.1.

Note that stratified trees are always also square-free.

Example 3 (Example 2 continued). *Let S, E and L be three binary random variables with a strictly positive joint probability mass function*

$$p_\theta(s, e, l) \;=\; \theta(s, e)\theta(s, e, l) \tag{2.14}$$

for all $(s, e, l) \in \{0, 1\}^3$. Here, S represents the social status of a child, E the economic background and L the number of life events. We use the numerical values 1 for 'high' and 0 for 'low'. Then the staged tree $(\mathcal{T}, \Theta_\mathcal{T})$ in Fig. 1 is (S, E, L)-compatible and stratified. The staging of $(\mathcal{T}, \Theta_\mathcal{T})$ is equivalent to the conditional independence assumption $E \perp\!\!\!\perp L \mid S$. This also can be represented by the three Markov-equivalent *DAGs $E \to S \to L$, $E \leftarrow S \leftarrow L$ and $E \leftarrow S \to L$ [1]. Note that both the staged tree and DAG model representations assert that $\theta(s, e, l) = \theta(s, e', l)$ for $e \neq e'$ and $s, l = 0, 1$, as in (2.13). We will henceforth write $\theta(s, l) = \theta(s, e, l)$ for primitive probabilities belonging to vertices in those stages.*

3. An algebraic characterization of staged tree models. Our analysis in this section starts by characterizing the subclass $[\mathcal{T}, \Theta_\mathcal{T}]^c \subseteq [\mathcal{T}, \Theta_\mathcal{T}]$ of those statistically equivalent probability trees which share the same associated polynomial c, as defined below. Substitution operations on c will then enable us to generalize from $[\mathcal{T}, \Theta_\mathcal{T}]^c$ to $[\mathcal{T}, \Theta_\mathcal{T}]$. Theorem 1 finds a complete characterization of $[\mathcal{T}, \Theta_\mathcal{T}]$ for any staged tree model $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}$.

3.1. *Stage ideals and polynomials.* Some of our results are expressed algebraically, so we first recall some notation from [7]. This enables us to move from a setting where parameters are place holders for as yet undetermined values to a setting where parameters are elements of some formal algebraic structure, see e.g. [9, 24] and [8, 15].

Let $S = \{s_1, \ldots, s_d\}$, $d \in \mathbb{N}$, be a finite set of indeterminates. A *(real) polynomial ring* $(\mathbb{R}_d[S], +, \cdot)$ is a commutative ring whose elements are formal polynomials $f = \sum_{i\in[n]} r_i s_1^{\alpha_{i,1}} s_2^{\alpha_{i,2}} \cdots s_d^{\alpha_{i,d}}$ with coefficients $r_i \in \mathbb{R}$ and exponents $\alpha_i \in \mathbb{N}_0^d$, for $i \in [n]$ and $n \in \mathbb{N}$. Then $f = g \in \mathbb{R}_d[S]$ in a formal sense if and only if $f = g : \ \mathbb{R}^d \to \mathbb{R}$ coincide as functions. The map between an interpretation of a polynomial as an element of a ring structure and as a function is called the *evaluation homomorphism.*

An *ideal* $I \subseteq \mathbb{R}_d[S]$ is a set of polynomials where $f \cdot g \in I$ for all $f \in \mathbb{R}_d[S]$ and $g \in I$, and where $(I, +)$ is a subgroup of $(\mathbb{R}_d[S], +)$. In particular,

$$\langle g_1, \ldots, g_k\rangle = \Big\{\sum_{i\in[k]} f_i g_i \mid f_i \in \mathbb{R}_d[S],\ i \in [k]\Big\} \subseteq \mathbb{R}_d[S] \tag{3.1}$$

defines an ideal *generated by* polynomials $g_1, \ldots, g_k \in \mathbb{R}_d[S]$. The set of zeros of these

$$V(I) = \big\{x \in \mathbb{R}^d \mid g(x) = 0 \text{ for all } g \in I\big\} \tag{3.2}$$

is called a *variety.*

The relation $\sim$ on $\mathbb{R}_d[S] \times \mathbb{R}_d[S]$ which is defined by $g \sim h$ if and only if $g - h \in I$, induces a ring of residue classes $\mathbb{R}_d[S]/_\sim = \mathbb{R}_d[S]/I$, called a *factor ring.* Note that by construction, the ideal $I \subseteq \mathbb{R}_d[S]$ is the residue class of all zeros in $\mathbb{R}_d[S]/I$.

The properties of ideals and varieties have already been widely and successfully used to capture and exploit the structure of (graphical) statistical models, see e.g. [9, 10, 13, 14]. In Section 2.2 we provided a graphical and geometric interpretation of the colouring in a given staged tree. This correspondence is expressed in terms of ideals and polynomial rings below.

Let $\mathbb{P}_\Psi$ be a discrete parametric statistical model with monomial parametrization $\Psi : \theta \mapsto \big(\theta_1^{\alpha_{i,1}} \cdots \theta_d^{\alpha_{i,d}} \mid \alpha_i \in \mathbb{N}_0^d, i \in [n]\big)$ as in Section 2.1. In an algebraic framework, we call the atomic probabilities in that vector *atomic monomials.* By abuse of notation, we always let $\mathbb{R}_d[\theta_1, \ldots, \theta_d] = \mathbb{R}_d[\Theta]$ denote the real polynomial ring in all indeterminates in a model parametrized by $\Psi : \Theta \to \mathbb{P}_\Psi$. In tree models $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}$, that ring is denoted $\mathbb{R}[\Theta_\mathcal{T}] = \mathbb{R}_d[\theta(e) \mid e \in E]$ with indeterminates given by primitive probabilities, for some $d \in \mathbb{N}$. This is again particular to one parametrization $\Psi_\mathcal{T}$ and one representation $(\mathcal{T}, \Theta_\mathcal{T})$, $\mathcal{T} = (V, E)$.

As in Section 2.2, let $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})} \subseteq \mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})_{\text{sat}}}$. The map linking $(\mathcal{T}, \Theta_\mathcal{T})_{\text{sat}}$ to $(\mathcal{T}, \Theta_\mathcal{T})$ is now the ring homomorphism

$$\begin{aligned} \Phi_\mathcal{T} : \quad & \mathbb{R}[\Theta_{\mathcal{T}_{\text{sat}}}] \to \mathbb{R}[\Theta_\mathcal{T}], \\ & \theta(e) \mapsto \theta(e') \text{ whenever } e \in E(v),\ e' = E(u) \text{ and } v \in u \in U_\mathcal{T} \end{aligned} \tag{3.3}$$

in analogy to the projection in (2.12). The kernel $\ker(\Phi_\mathcal{T})$ equals the ideal

$$I_\mathcal{T} = \langle \theta(e) - \theta(e') \mid e \in E(v),\ e' \in E(u) \text{ and } v \in u \in U_\mathcal{T} \rangle. \tag{3.4}$$

Note that the degree 1 binomials generating $I_\mathcal{T}$ capture the componentwise equations $\theta_v = \theta_{v'}$ whenever $v, v' \in u$. We call $I_\mathcal{T}$ the *stage ideal* of $(\mathcal{T}, \Theta_\mathcal{T})$. The *isomorphism theorem* [7] then implies that the factor ring $\mathbb{R}[\Theta_{\mathcal{T}_{\text{sat}}}]/I_\mathcal{T}$ is isomorphic to the polynomial ring $\mathbb{R}[\Theta_\mathcal{T}]$. This can be interpreted as first drawing a saturated tree and then embedding stage information yields the same model representation as the one obtained by directly drawing a staged tree.

The variety belonging to the stage ideal equals

$$V(I_\mathcal{T}) = \big\{ \theta(e) = \theta(e') \mid e \in E(v),\ e' \in E(u) \text{ and } v \in u \in U_\mathcal{T} \big\}. \tag{3.5}$$

Note that $V(I_\mathcal{T})$ equals the space spanned by all indeterminates in $\Theta_\mathcal{T}$, as a subset of the one spanned by all indeterminates in $\Theta_{\mathcal{T}_{\text{sat}}}$. Thus, the parameter set of a staged tree can be identified from its saturated version together with the stage ideal.

The equations in primitive probabilities constraining a staged tree can be easily translated into polynomial constraints on atomic probabilities $\pi_i = P_\theta(\omega_i)$, $i \in [n]$. These equations specify an ideal in a ring $\mathbb{R}_n[\pi_1, \ldots, \pi_n]$ which is the same across all graphical representations of a staged tree model $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}$. Then $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})} \subseteq \Delta_{n-1}$ can be characterized as the variety of this *ideal of model invariants*, and is thus an *algebraic statistical model* [10].

We illustrate the relationship between these two representations below.

Example 4 (Examples 2 and 3 continued). *The stage ideal of the staged tree $(\mathcal{T}, \Theta_\mathcal{T})$ in Fig. 1 equals*

$$I_\mathcal{T} = \langle \theta(s, e, l) - \theta(s, e', l) \mid e \neq e', s, l = 0, 1 \rangle \tag{3.6}$$

in the polynomial ring $\mathbb{R}[\Theta_{\mathcal{T}_{\text{sat}}}] = \mathbb{R}_{12}[\theta(s, e), \theta(s, e, l) \mid s, e, l = 0, 1]$ of the saturated model. As above, the factor ring $\mathbb{R}[\Theta_{\mathcal{T}_{\text{sat}}}]/I_\mathcal{T}$ is isomorphic to the polynomial ring $\mathbb{R}[\Theta_\mathcal{T}] = \mathbb{R}_8[\theta(s, e), \theta(s, l) \mid s, l = 0, 1]$ of the staged tree.

In [9, 13], the authors algebraically characterize the conditional independence assumption in this model using the cross-product differences *in*

$$I_p = \langle p_{000}p_{101} - p_{001}p_{100},\ p_{010}p_{111} - p_{011}p_{110}\rangle \tag{3.7}$$

in a polynomial ring $\mathbb{R}[p] = \mathbb{R}_8[p_{sel} \mid s, e, l = 0, 1]$ *with* $p_{sel} = p_\theta(s, e, l)$ *for* $s, e, l \in \{0, 1\}$*, from (2.14). These constraints on the joint probabilities are the same for every DAG representation of the model and hence index the Markov-equivalence class. They could not be read from a staged tree.*

Note that both $I_\mathcal{T}$ *and* $V(I_\mathcal{T})$ *are much simpler objects than* I_p *and* $V(I_p)$*. A reparametrization between the two model representations is given by the ring homomorphism* $\Phi : \mathbb{R}[p] \to \mathbb{R}[\Theta_\mathcal{T}]$, $p_{sel} \mapsto \theta(s, e)\theta(s, l)$*. The inverse of this map is rational, calculated using the law of total probability,*

$$\Phi^{-1} : \quad \theta(s, e) \mapsto p_{se0} + p_{se1},\ \theta(s, l) \mapsto \frac{p_{s0l} + p_{s1l}}{p_{s00} + p_{s01} + p_{s10} + p_{s11}}. \tag{3.8}$$

Thus, whilst Φ *is an invertible function and both ideals capture in this sense the same information,* Φ *is not a polynomial ring isomorphism, and* $I_\mathcal{T}$ *and* I_p *are very different objects.*

In this paper we develop an alternative but rather different algebraic approach to the standard one of using ideals and varieties, which is more intuitive for staged trees. By using the polynomial defined below we are able to recover constructively all possible graph representations of the same model, a process not possible using only stage ideals or ideals of model invariants. In particular, when treating primitive probabilities as formal indeterminates, we can easily move from one representation of a model to the next, ignoring sum-to-1 conditions (and thus the normalization of the underlying probability mass function) which come for free in tree models: see Lemma 1.

3.2. *Polynomially equivalent staged trees and the swap operator.* Following [8, 15, 25], we define:

Definition 3 (Interpolating polynomial). *Let* $\mathbb{P}_\Psi \subseteq \Delta^\circ_{n-1}$ *be a discrete statistical model with underlying discrete space* Ω *and monomial parametrization* $\Psi : \theta \mapsto \boldsymbol{p}_\theta \in \mathbb{P}_\Psi$*,* $\theta \in \Theta$*. A* network polynomial *is of the form*

$$c_{g,\Psi}(\theta) = \sum_{\omega \in \Omega} g(\omega) p_\theta(\omega) \tag{3.9}$$

where $g : \Omega \to \mathbb{R}$ *is a polynomial. If* $g = 1$*, then the formal sum* $c = c_{1,\Psi}$ *is called an* interpolating polynomial *for* $\mathbb{P}_\Psi$*.*

Note that in [8], $g = \mathbb{1}_A$ is an indicator of events $A \in \sigma(\Omega)$ and c_g is called the network polynomial of a discrete Bayesian network. By [36], indicator functions are indeed polynomials. In other publications we have already demonstrated the efficacy of using network polynomials in calculating marginal and conditional probabilities in staged tree models [15] and for sensitivity analysis in models with a multilinear parametrization [22].

If the parametric model in the definition above is a staged tree model $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}$ then we also write $c_\mathcal{T} = \sum_{\lambda\in\Lambda(\mathcal{T})} \pi_{\theta,\mathcal{T}}(\lambda) \in \mathbb{R}[\Theta_\mathcal{T}]$ for the interpolating polynomial $c_{1,\Psi_\mathcal{T}} = c_\mathcal{T}$ of a specific representation $(\mathcal{T},\Theta_\mathcal{T}) \in [\mathcal{T},\Theta_\mathcal{T}]$. Note that $(\mathcal{T},\Theta_\mathcal{T})$ is square-free if and only if its parametrization $\Psi_\mathcal{T}$ is multilinear, so if and only if $c_\mathcal{T}$ is linear in every indeterminate in $\mathbb{R}[\Theta_\mathcal{T}]$.

We are now able to examine a set of properties of the interpolating polynomial as an object in a polynomial ring. We first define a class of tree representations for which the interpolating polynomial is formally invariant.

Definition 4 (Polynomial equivalence). *Let $(\mathcal{T},\Theta_\mathcal{T})$, $(\mathcal{S},\Theta_\mathcal{S})$ be two staged trees with the same underlying space Ω. The trees are* polynomially equivalent *if and only if they have the same sets of primitive probabilities and their network polynomials coincide formally $c_{g,\mathcal{S}} = c_{g,\mathcal{T}} \in \mathbb{R}[\Theta_\mathcal{T}]$ for all $g \in \mathbb{R}[\Theta_\mathcal{T}]$.*

Henceforth, the symbol $[\mathcal{T},\Theta_\mathcal{T}]^c$ denotes a class of polynomially equivalent staged trees sharing the same interpolating polynomial c. Example 5 will present staged trees which are polynomially equivalent to the one analyzed in Examples 2 to 4.

When two network polynomials are equal for any polynomial g, they are also termwise equal and the corresponding atomic probabilities can be identified. So we have the following:

Lemma 2. *Polynomial equivalence implies statistical equivalence.*

In general, polynomial equivalence is not necessary for statistical equivalence, see e.g. Example 1. However, Examples 6 and 7 present interesting cases where necessity holds. Note that by Lemma 1, the composition of floret parameter vectors between polynomially equivalent trees can differ. We will thus think of polynomially equivalent trees as a set of graphical representations of a model which share the same algebraic representation in terms of potentials (see page 8) with different normalizations.

By Definition 3, the interpolating polynomial $c_\mathcal{T}$ of a staged tree is simply a sum over all atomic monomials, calculated as the product of all edge labels along a root-to-leaf path in $\mathcal{T} = (V, E)$. Centrally, the graph also yields a way to parenthesize the interpolating polynomial: For every floret $\mathcal{F}_v$ where $v \in V$ is the parent of a leaf, we sum all components of its parameter vector θ_v and multiply the result by its parent label $\theta(\mathrm{pa}(v), v)$. We then sum the result over the parent's labels $\theta_{\mathrm{pa}(v)}$. By repeating this until all floret parameter vectors are summed and $\mathrm{pa}(v) = v_0$, the interpolating polynomial can then be written in terms of the nested factorization

$$c_{\mathcal{T}}(\theta) = \sum_{v_1 \in \mathrm{ch}(v_0)} \theta(v_0, v_1) \Big(\sum_{v_2 \in \mathrm{ch}(v_1)} \theta(v_1, v_2) \; \dots \; \Big(\sum_{v_k \in \mathrm{ch}(v_{k-1})} \theta(v_{k-1}, v_k) \Big) \Big). \tag{3.10}$$

More generally,

Definition 5 (Tree compatibility). *Let $\mathbb{P}_\Psi$ be a discrete parametric model with monomial parametrization and associated polynomial ring $\mathbb{R}_d[\Theta]$. We call any polynomial $c \in \mathbb{R}_d[\Theta]$* tree compatible *if it admits a representation of the form*

$$c(\theta) = \sum_{\theta_1 \in A_1} \theta_1 \Big(\sum_{\theta_2 \in A_2(\theta_1)} \theta_2 \Big(\sum_{\theta_3 \in A_3(\theta_2)} \theta_3 \; \dots \; \Big(\sum_{\theta_k \in A_k(\theta_{k-1})} \theta_k \Big) \Big) \Big) \tag{3.11}$$

where every $\#A_1, \#A_j(\theta_{j-1}) \geq 2$ for $j \in [k]$, $k \in \mathbb{N}$. We write $s(c(\theta))$ for one fixed order of summation of the terms in $c(\theta)$ as above, and call this a tree-compatible factorization.

The interpolating polynomial $c_{\mathcal{T}}(\theta) = \sum_{x \in \mathbb{X}} \pi_{\theta, \mathcal{T}}(\iota_{\mathcal{T}}(x))$ of every X-compatible tree $(\mathcal{T}, \Theta_{\mathcal{T}})$ admits an explicit tree-compatible factorization

$$c_{\mathcal{T}}(\theta) = \sum_{x_{A_1} \in \mathbb{X}_{A_1}} \theta(x_{A_1}) \Big(\sum_{x_{A'_2} \in \mathbb{X}_{A'_2}} \theta(x_{A_2}) \Big(\sum_{x_{A'_3} \in \mathbb{X}_{A'_3}} \theta(x_{A_3}) \dots \Big(\sum_{x_{A'_k} \in \mathbb{X}_{A'_k}} \theta(x_{A_k}) \Big) \Big) \Big) \tag{3.12}$$

where the index sets A_i are as in (2.11) and $A'_i = A_i \setminus A^{i-1}$, $i \geq 2$. Note here that by Proposition 1 below, the embedding in (2.11) can be recovered from (3.12). Every factorization like the one above together with its corresponding embedding then interpret the underlying monomial parametrization $\Psi_{\mathcal{T}}$ in a non-commutative way. Different tree-compatible factorizations of the interpolating polynomial thus correspond to different normalizations of the underlying probability mass function: see also Example 5 below.

The factorization (3.12) provides a very efficient way to compute joint probabilities from marginals in a BN model [20] and comes for free when representing a discrete BN by a stratified tree.

An important aspect of the above result is that it is reversible: not only can we easily read a polynomial from a tree graph, but we can also construct a tree graph from a tree-compatible factorization. In addition, all polynomially equivalent staged trees arise from a tree-compatible reordering of a given summation. Each of these gives a different representations within the same statistical equivalence class.

Proposition 1. *Let $\mathbb{P}_\Psi$ be a discrete parametric model with multilinear monomial parametrization Ψ and interpolating polynomial $c = c_{1,\Psi} \in \mathbb{R}_d[\Theta]$. Then there exists a probability tree $(\mathcal{T}, \Theta_\mathcal{T})$ with $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})} = \mathbb{P}_\Psi$ if and only if c is tree compatible.*

The map $\mathfrak{c} : \{s(c(\theta)) \mid s\} \to [\mathcal{T}, \Theta_\mathcal{T}]^c$, $s(c(\theta)) \mapsto (\mathcal{T}, \Theta_\mathcal{T})$ is bijective.

The above proposition provides us with a powerful tool to decide if a parametric model can be represented by a probability tree. This representation is a staged tree only if all constraints on the model are of the form $A_{i+1}(\theta_i) = A_{j+1}(\theta_j)$ for some $i \neq j$ in the notation of Definition 5. Note that because Ψ above is always multilinear, it follows that this tree is also square-free. This constraint is necessary in Proposition 1 because only then is $\pi_{\theta,\mathcal{T}} : \Lambda(\mathcal{T}) \to \mathbb{R}[\Theta_\mathcal{T}]$ a formally injective function, implying we can uniquely identify atoms (or root-to-leaf paths) with their atomic monomials.

Now, the result above provokes two natural questions. First, how can we decide whether or not a given interpolating polynomial $c \in \mathbb{R}_d[\Theta]$ is tree compatible? We can answer this question based on calculating the greatest common divisor of certain terms in c and performing a sequence of ideal membership tests for projections of c onto subrings of $\mathbb{R}_d[\Theta]$.

The second question is: how do we infer all the possible orders of bracketing of a tree-compatible interpolating polynomial $c_\mathcal{T}$? If we are able to do this, then, using the map $\mathfrak{c}$ in Proposition 1 and the construction outlined in the proof, we can obtain all tree representations in $[\mathcal{T}, \Theta_\mathcal{T}]^c$. Either by a direct elicitation of a model or by applying a model selection methodology, typically there is a particular staged tree representation whose interpretation we then need to draw out. A characterization of all its polynomially (and hence statistically) equivalent representations will then be vital if this interpretation is to be given unambiguously. So we next study how to find different tree-compatible factorizations of a given polynomial.

Clearly, a transformation between two tree-compatible factorizations of an interpolating polynomial is an application of the distributive property of addition and multiplication in the ring $(\mathbb{R}[\Theta_\mathcal{T}], +, \cdot)$. We show below that a map between polynomially equivalent staged trees can then be characterized by a finite number of corresponding intuitive graph transformations.

Let $\mathcal{T} = (V, E)$. Let $\mathcal{T}' = (V', E')$ where $(V', E') \subseteq (V, E)$ be an event tree with induced edge labels $\theta(e') = \theta(e)$ for $e' = e \in E' \cap E$. We henceforth call $(\mathcal{T}, \Theta_\mathcal{T})'$ a *(probability) subtree* and write $\mathcal{T}' \subseteq \mathcal{T}$ and $(\mathcal{T}, \Theta_\mathcal{T})' \subseteq (\mathcal{T}, \Theta_\mathcal{T})$. Call a probability subtree an *adjacency subtree* if it is of the following form: $\mathcal{T}'$ has two levels and all children of its root $v_0' \in V'$ are in the same stage $u \in \tilde{U}_\mathcal{T}$ in $\mathcal{T} \supseteq \mathcal{T}'$. All florets $\mathcal{F}_{v'}$ with $v' \in \mathrm{ch}(v_0')$ are induced by $\mathcal{T}$, in the sense that all edges in $E(v') \subseteq E$ are also in E'. If $v_0' \in u'$ is in a non-trivial stage $u' \in \tilde{U}_\mathcal{T}$ in $\mathcal{T}$, then also the root-floret $\mathcal{F}_{v_0'}$ of $\mathcal{T}'$ is induced by $\mathcal{T}$.

The interpolating polynomial of an adjacency subtree $(\mathcal{T}, \Theta_\mathcal{T})'$ equals

$$c_{\mathcal{T}'}(\theta) = \sum_{e' \in E(v_0')} \theta(e') \Big(\sum_{e \in E(u)} \theta(e) \Big) = \sum_{e \in E(u)} \theta(e) \Big(\sum_{e' \in E(v_0')} \theta(e') \Big) \tag{3.13}$$

with $E(u) = E(v)$ for one $v \in u$. Thus, by Proposition 1, there is precisely one *staged* tree $(\mathcal{T}, \Theta_\mathcal{T})'_u$ which is polynomially equivalent to $(\mathcal{T}, \Theta_\mathcal{T})'$: this is the one given by the second tree-compatible factorization in (3.13). Also, $(\mathcal{T}, \Theta_\mathcal{T})'_u$ is a subtree of the tree $(\mathcal{T}, \Theta_\mathcal{T})_u$ which is polynomially equivalent to $(\mathcal{T}, \Theta_\mathcal{T})$ and coincides with that tree everywhere except on $(\mathcal{T}, \Theta_\mathcal{T})'$.

Definition 6 (Swap). *Let $(\mathcal{T}, \Theta_\mathcal{T})$ be a staged tree, $(\mathcal{T}, \Theta_\mathcal{T})' \subseteq (\mathcal{T}, \Theta_\mathcal{T})$ an adjacency subtree with stage $u \in \tilde{U}_\mathcal{T}$. Denote by $(\mathcal{T}, \Theta_\mathcal{T})'_u$ the staged tree which is polynomially equivalent to $(\mathcal{T}, \Theta_\mathcal{T})'$, and $(\mathcal{T}, \Theta_\mathcal{T})'_u \subseteq (\mathcal{T}, \Theta_\mathcal{T})_u$. We then call the map $\mathfrak{s} :\ (\mathcal{T}, \Theta_\mathcal{T}) \mapsto (\mathcal{T}, \Theta_\mathcal{T})_u$ a* swap.

Figure 2 illustrates an adjacency subtree and a swap. We can see there that this operation does indeed 'swap' the order of edges before and after the stage u. By [35], edge-centred events $\Lambda(e)$, $e = (v_0', v) \in E'$, on the first level of an adjacency subtree are independent of those $\Lambda(e')$, $e' = (v, v') \in E'$, on the second level. Our very plausible discovery is that for these independent events the order $\Lambda(e) \prec \Lambda(e')$ is reversible within a statistical equivalence class, using the swap operator. This result is used in a causal analysis in Section 4.

We henceforth call a composition of swaps for which floret parameter vectors are invariant a *floret-swap*. For instance, the swap in Fig. 2 is not a floret-swap because the root-vector $(\theta_1, \theta_2, \theta_3) \in \Theta_\mathcal{T}$ is not an element of $\Theta_\mathcal{S}$, and conversely $(\theta_1, \theta_4, \theta_5) \in \Theta_\mathcal{S} \setminus \Theta_\mathcal{T}$. Importantly, $(\mathcal{T}, \Theta_\mathcal{T})$ and $(\mathcal{S}, \Theta_\mathcal{S})$ thus have different local sum-to-1 conditions on their primitive probabilities. By Lemma 1 and 2, both are still representations of the same model. So even if the numerical value of say $\theta_1 = \theta(e_1)$ is different in $(\mathcal{T}, \Theta_\mathcal{T})$ and $(\mathcal{S}, \Theta_\mathcal{S})$—we

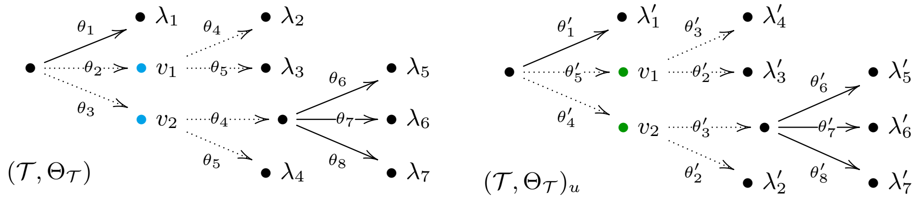

FIGURE 2. *Two polynomially equivalent staged trees with the same indeterminates $\theta_i = \theta'_i$ and identified root-to-leaf paths $\lambda_i = \iota_\mathcal{T} \circ \iota_\mathcal{S}^{-1}(\lambda'_i)$ for all $i \in [8]$. The adjacency subtrees $(\mathcal{T}, \Theta_\mathcal{T})' \subseteq (\mathcal{T}, \Theta_\mathcal{T})$ and $(\mathcal{T}, \Theta_\mathcal{T})'_u \subseteq (\mathcal{T}, \Theta_\mathcal{T})_u$ are depicted by dotted lines and share the stage $u = \{v_1, v_2\}$. The map $\mathfrak{s} :\ (\mathcal{T}, \Theta_\mathcal{T}) \mapsto (\mathcal{T}, \Theta_\mathcal{T})_u$ is a swap. See page 17.*

have indicated this using lables $\theta_1 = \theta'_1$ in Fig. 2—via a renormalization it is still the probability of the event $\iota_\mathcal{T}^{-1}(\Lambda(e_1)) \in \sigma(\Omega)$. The meaning of this parameter is thus unchanged and can be identified across different graphs.

A composition of swaps which permute two levels of a tree is called a *level-swap*. Level-swaps map stratified trees into stratified trees: see again Section 4 for an application.

We can now obtain the following result, which enables us to both graphically and algebraically move around a class of polynomially equivalent trees.

PROPOSITION 2. *Two square-free staged trees $(\mathcal{T}, \Theta_\mathcal{T})$, $(\mathcal{S}, \Theta_\mathcal{S})$ are polynomially equivalent if and only if there exists a finite composition of swaps $\mathfrak{s}_1, \ldots, \mathfrak{s}_l$, $l \in \mathbb{N}$, for which*

$$\mathfrak{s} = \mathfrak{s}_l \circ \mathfrak{s}_{l-1} \circ \ldots \circ \mathfrak{s}_1 :\ (\mathcal{T}, \Theta_\mathcal{T}) \mapsto (\mathcal{S}, \Theta_\mathcal{S}). \tag{3.14}$$

Thus, the polynomial equivalence classes of staged trees can be fully characterized by the local graph transformations given by swaps. Note that this operator is a close tree analogue of an *arc reversal* in BN models. These, just like swaps, allow one to traverse the class of all graphical representations of the same model [1], while renormalizing (but not marginalizing) the associated probability mass function: see the example below.

EXAMPLE 5 (Examples 2 to 4 continued). *Consider again the staged tree $(\mathcal{T}, \Theta_\mathcal{T})$ in Fig. 1. Its interpolating polynomial $c_\mathcal{T}$ admits the following four tree-compatible factorizations, in the notation of Example 3.*

$$c_\mathcal{T}(\theta) = \sum_{s,e,l=0,1} p_\theta(s,e,l) = \sum_{s,e=0,1} \theta(s,e)\Big(\sum_{l=0,1} \theta(s,l)\Big) \tag{3.15a}$$

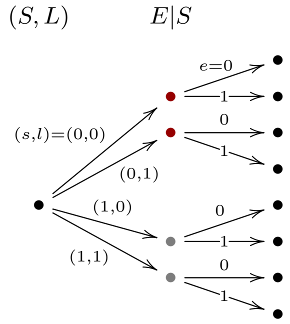

FIGURE 3. *An 'arc reversal' staged tree $(\mathcal{S},\Theta_{\mathcal{S}})_1$ which is polynomially equivalent to $(\mathcal{T},\Theta_{\mathcal{T}})$ from Example 2, Fig. 1. See Example 5.*

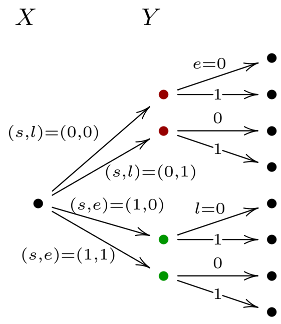

FIGURE 4. *A 'twist' staged tree $(\mathcal{S},\Theta_{\mathcal{S}})_2$ from Example 5 which is polynomially equivalent to the ones in Figs. 1 and 4.*

$$= \sum_{s,l=0,1} \theta(s,l)\Big(\sum_{e=0,1}\theta(s,e)\Big) \tag{3.15b}$$

$$= \sum_{l=0,1} \theta(0,l)\Big(\sum_{e=0,1}\theta(0,e)\Big) + \sum_{e=0,1}\theta(1,e)\Big(\sum_{l=0,1}\theta(1,l)\Big) \tag{3.15c}$$

$$= \sum_{l=0,1} \theta(1,l)\Big(\sum_{e=0,1}\theta(1,e)\Big) + \sum_{e=0,1}\theta(0,e)\Big(\sum_{l=0,1}\theta(0,l)\Big). \tag{3.15d}$$

We denote the staged trees corresponding to (3.15a) to (3.15d) by $(\mathcal{T},\Theta_{\mathcal{T}})$, $(\mathcal{S},\Theta_{\mathcal{S}})_1$, $(\mathcal{S},\Theta_{\mathcal{S}})_2$ and $(\mathcal{S},\Theta_{\mathcal{S}})_3$, respectively.

From Example 4, $(\mathcal{T},\Theta_{\mathcal{T}})$ gives an alternative representation of a BN, and we can label its levels by the random variables (S,E) and $L|S$. Now, $\mathfrak{s}_1 : (\mathcal{T},\Theta_{\mathcal{T}}) \mapsto (\mathcal{S},\Theta_{\mathcal{S}})_1$ is a level-swap and $(\mathcal{S},\Theta_{\mathcal{S}})_1$ in Fig. 3 represents the joint variable (S,L) first and $E|S$ last. Since $(\mathcal{T},\Theta_{\mathcal{T}})$ and $(\mathcal{S},\Theta_{\mathcal{S}})_1$ are in the same polynomial equivalence class, we can deduce here that $[\mathcal{T},\Theta_{\mathcal{T}}]^c$ is sufficiently rich to contain tree representations which renormalize $\pi_{\theta,\mathcal{T}}$ to $\pi_{\theta,\mathcal{S}}(\iota_{\mathcal{S}}(s,e,l)) = \theta(s,l)\theta(s,e)$. See also Example 7.

Note that, unlike $(\mathcal{S},\Theta_{\mathcal{S}})_1$, the staged tree $(\mathcal{S},\Theta_{\mathcal{S}})_2 = \mathfrak{s}_2(\mathcal{T},\Theta_{\mathcal{T}})$ in Fig. 4 is not (S,E,L)-compatible. This tree now belongs to a different DAG $X \to Y$ where

$$X = \begin{cases} (S,L) & \text{if } S=0 \\ (S,E) & \text{if } S=1 \end{cases} \quad \text{and} \quad Y = \begin{cases} E|L & \text{if } S=0 \\ L|E & \text{if } S=1. \end{cases}$$

We call such a transformation a twist. *$(\mathcal{S},\Theta_{\mathcal{S}})_3$ is also a twist of $(\mathcal{T},\Theta_{\mathcal{T}})$.*

The example above provides a very simple illustration of how the statistical equivalence classes of a staged tree (or CEG) are so much larger than those of the BN. It also demonstrates how staged trees can implicitly generate relationships between new random variables, constructed as functions of the original ones: possibly useful in later interpretative analysis. A more detailed discussion of this process is given in Section 4.1. See also [6, 30, 35].

Note that by Example 1, the polynomial equivalence class of a staged tree can be a proper subclass of its statistical equivalence class. In fact, because saturated probability trees have no adjacency subtrees, we find:

Corollary 1. *Let $(\mathcal{T}, \Theta_{\mathcal{T}})$ be a saturated probability tree with interpolating polynomial $c = c_{\mathcal{T}}$. Then $\#[\mathcal{T}, \Theta_{\mathcal{T}}]^c = 1$.*

We will need this rigidity of polynomial equivalence in cases where we want to prevent a renormalization of the underlying probability mass function or to maintain a fixed order of edge-centred events, as in Example 6 below. The following subsection outlines how to impose or avoid it when necessary.

3.3. *Statistically equivalent staged trees and the resize operator.* We now extend the characterization of polynomial equivalence classes $[\mathcal{T}, \Theta_{\mathcal{T}}]^c$, with $c \in \mathbb{R}[\Theta_{\mathcal{T}}]$, to classes $[\mathcal{T}, \Theta_{\mathcal{T}}]$ of statistically equivalent staged trees. This extension is based on reparametrizations between the associated polynomial rings $\mathbb{R}[\Theta_{\mathcal{T}}]$ and $\mathbb{R}[\Theta_{\mathcal{S}}]$ for $(\mathcal{T}, \Theta_{\mathcal{T}})$, $(\mathcal{S}, \Theta_{\mathcal{S}}) \in [\mathcal{T}, \Theta_{\mathcal{T}}]$.

Lemma 3. *Let $\mathbb{P}_{(\mathcal{T},\Theta_{\mathcal{T}})}$ be a staged tree model with an underlying probability space $(\Omega, \sigma(\Omega), P_\theta)$ and let $(\mathcal{T}, \Theta_{\mathcal{T}}) \in [\mathcal{T}, \Theta_{\mathcal{T}}]$ be one representation with embedding $\iota_{\mathcal{T}} : \Omega \to \Lambda(\mathcal{T})$. Then for any $A \in \sigma(\mathcal{T})$ and $g = \mathbb{1}_A$, the network polynomial c_g from (3.9) is a map*

$$\begin{aligned} c_{\mathbb{1}_\cdot,\mathcal{T}}: \quad \sigma(\mathcal{T}) \times \bigtimes_{v\in V} \Delta^\circ_{\#E(v)-1} &\to [0,1] \\ (A, \theta) &\mapsto P_\theta(\iota_{\mathcal{T}}^{-1}(A)). \end{aligned} \tag{3.16}$$

Moreover, for any statistically equivalent $(\mathcal{S}, \Theta_{\mathcal{S}}) \in [\mathcal{T}, \Theta_{\mathcal{T}}]$, the polynomials $c_{\mathbb{1}_\cdot,\mathcal{T}} = c_{\mathbb{1}_\cdot,\mathcal{S}}$ are equal as functions.

Thus, while the interpolating polynomial as a formal polynomial is unique to a polynomial equivalence class, as a function it is unique to the whole statistical equivalence class. Below, we define a second operator which uses this result constructively.

We call the pair $(\mathcal{T}, \Theta_{\mathcal{T}})' \subseteq (\mathcal{T}, \Theta_{\mathcal{T}})$ a *(probability) subgraph* if it is a probability subtree whose root might have only one emanating edge (not two as required in an event tree). It is easily deduced that every staged tree can be characterized by a set of probability subgraphs. For efficient computation it is often helpful to replace these subgraphs by florets.

Definition 7 (Resize). *Let $(\mathcal{T}, \Theta_\mathcal{T})$ be a staged tree and let*

$$\mathfrak{r}: \ (\mathcal{T}, \Theta_\mathcal{T}) \mapsto (\mathcal{S}, \Theta_\mathcal{S}) \tag{3.17}$$

denote the map which transforms a probability subgraph $(\mathcal{T}, \Theta_\mathcal{T})' \subseteq (\mathcal{T}, \Theta_\mathcal{T})$ into a probability tree $(\mathcal{F}, \{\theta_\mathcal{F}\})$ whose graph is a floret with parameter vector $\theta_\mathcal{F} = \big(\pi_{\theta,\mathcal{T}'}(\lambda') \mid \lambda' \in \Lambda(\mathcal{T}')\big)$. Then $\mathfrak{r}$ maps the whole tree $(\mathcal{T}, \Theta_\mathcal{T})$ to a probability tree $(\mathcal{S}, \Theta_\mathcal{S})$ which has that floret as a probability subtree and is identical to $(\mathcal{T}, \Theta_\mathcal{T})$ otherwise.

We call $\mathfrak{r}$ and its inverse $\mathfrak{r}^{-1}$ naive resize *operators, and a* resize *if $(\mathcal{S}, \Theta_\mathcal{S})$ is a staged tree.*

In terms of the atomic monomials, a naive resize performs a substitution of products of primitive probabilities into degree 1 monomials. By Lemma 3, atomic probabilities are invariant under this operation. The saturated tree and its floret representation are again a natural example of this, see Example 1. Thus, every naively resized $\mathfrak{r}(\mathcal{T}, \Theta_\mathcal{T})$ is clearly a probability tree representing the same model $\mathbb{P}_{(\mathcal{T},\Theta_\mathcal{T})}$ as $(\mathcal{T}, \Theta_\mathcal{T})$. The lemma below establishes various useful criteria under which $\mathfrak{r}$ is a well-defined map between two *staged* trees.

Lemma 4. *Let $(\mathcal{T}, \Theta_\mathcal{T})$ be a staged tree and $l \in \mathbb{N}$. A composition of naive resizes $\mathfrak{r} = \mathfrak{r}_l \circ \ldots \circ \mathfrak{r}_1$ applied to $(\mathcal{T}, \Theta_\mathcal{T})$ is a resize if one of the following conditions is fulfilled:*

a) $\mathfrak{r}$ only acts on saturated probability subgraphs.

b) $\mathfrak{r}$ only acts on probability subgraphs which are polynomially equivalent to each other and whose vertices are not in the same stage as vertices outside these subgraphs.

Note that case (a) in Lemma 4 enables us to contract uninformative, trivially staged subgraphs into florets. Case (b) enables us to directly identify atomic monomials of polynomially equivalent subgraphs rather than repeating stage equations edge by edge. Note that if these conditions are violated, then a naive resize can take us out of the statistical equivalence class of a staged tree, as illustrated below.

Example 6. *Consider a BN given by binary random variables X_1, X_2 and X_3, and a* collider *DAG $X_1 \to X_3 \leftarrow X_2$ [29]. We can represent this* *model by an X-compatible staged tree $(\mathcal{T}, \Theta_{\mathcal{T}})$ with embedding $(x_1, x_2, x_3) \overset{\iota_{\mathcal{T}}}{\mapsto} (e_1, e_2, e_3)$ where $e_i = e(x_{[i]})$, $i = 1, 2, 3$. Then because $X_1 \perp\!\!\!\perp X_2$, all children of the root are in the same stage. So the subtree with root-to-leaf paths (e_1, e_2) is an adjacency subtree. The primitive probabilities $\theta(e_3) = \theta(x_{\{1,2,3\}})$ however will be pairwise different for $x_{[3]} \in \mathbb{X}_{[3]}$ because X_3 is not (conditionally) independent of any of the other two variables.*

The polynomial equivalence class of such a staged tree enables us to level-swap X_1 and X_2, in the same way that we can permute these vertices in a DAG representation, and keeps X_3 fixed. This order of variables, 'X_1 and X_2 happen before X_3', can be interpreted as having a possible chronological or causal meaning [23]. See again Section 4.

Now, any naive resize operator on $(\mathcal{T}, \Theta_{\mathcal{T}})$ would substitute factors in an atomic monomial by terms of lower degree. As $\theta(e(x_1, x_2)) = \theta(e(x_1', x_2))$ for all $x_2 \in \mathbb{X}_2$, $x_1 \neq x_1' \in \mathbb{X}_1$, this information would then need to be captured by a set of non-linear equations, or cross-product differences as in Example 4. Any resize will thus not yield a staged tree model as defined in Section 2 but a tree model analogous to a context-specific BN, namely a graph together with some extra non-graphical information.

Now, for non-naive resizes, we obtain the following result which follows immediately from Lemma 3.

Lemma 5. *Let $(\mathcal{T}, \Theta_{\mathcal{T}})$ be a staged tree and $\mathfrak{r}$ a resize operator. Then $(\mathcal{T}, \Theta_{\mathcal{T}})$ and $(\mathcal{S}, \Theta_{\mathcal{S}}) = \mathfrak{r}(\mathcal{T}, \Theta_{\mathcal{T}})$ are statistically equivalent staged trees.*

The example below draws the link to analogue operations in BN models, and stresses that clever applications of resizes can enable us to restrict our analysis of a model to sufficiently expressive polynomial equivalence classes.

Example 7. *A* decomposable BN*, in accordance with [18], is a model with a probability mass function of the form*

$$p_\theta(x) = \prod_{j\in[k]} \theta(x_{C_j}) \quad \textit{for all } x \in \mathbb{X} \tag{3.18}$$

where C_i, $i \in [k]$, are cliques *of an underlying DAG D, that is maximally complete sets of vertices, and $B_j = C_j \cap C^j$, where $C^j = \bigcup_{i=1}^{j-1} C_i$ for $j = 2, 3, \ldots, k$, are* separators*. Then, $X_{Cj} \perp\!\!\!\perp X_{Cj-1 \setminus Bj} \mid X_{Bj}$ for all $j = 2, 3, \ldots, k$.*

The parametrization above is natural because there are no conditional independence constraints between variables within the same clique. Hence, inference is often made from a junction tree *[17] instead of from D, or in* *non-decomposable models from a* DAG of chain components *[19]. In analogy to a resized staged tree, these are also graphically more compact versions of the original DAG.*

Note that in Example 3 we represented the DAG $S \to E \to L$ by a staged tree with two levels which was actually based on a DAG $(S, E) \to L$. Using the swap operator and an extension of the argument presented in Example 5, we can show that the polynomial equivalence class of a staged tree $(\mathcal{T}, \Theta_{\mathcal{T}})$ with a clique-parametrization $\pi_{\theta,\mathcal{T}} = p_\theta$ as in (3.18) contains all X-compatible staged tree representations of the decomposable BN model.

The resize in conjunction with the swap operator now enables us to traverse the whole equivalence class of a given staged tree.

Theorem 1. *Two square-free staged trees $(\mathcal{T}, \Theta_{\mathcal{T}})$, $(\mathcal{S}, \Theta_{\mathcal{S}})$ are statistically equivalent if and only if there exists a map $\mathfrak{m} : (\mathcal{T}, \Theta_{\mathcal{T}}) \mapsto (\mathcal{S}, \Theta_{\mathcal{S}})$ which is a finite composition of resizes and swaps.*

In the last section we briefly demonstrate how the theorem above can be used to enhance the interpretation of a discovered staged tree model.

4. Analyzing the statistical equivalence class of a staged tree.

Feasible orders of events in statistically equivalent staged tree representations are most easily analyzed within a polynomial equivalence class. This is because, by Proposition 2, this class can be traversed using a swap $\mathfrak{s}$. In particular, if $(e, e') \in \Lambda(\mathcal{T}')$ is a root-to-leaf path in an adjacency subtree $(\mathcal{T}, \Theta_{\mathcal{T}})' \subseteq (\mathcal{T}, \Theta_{\mathcal{T}})$ then $(e', e) \in \Lambda(\mathcal{S}')$ is a subpath in the image $\mathfrak{s}(\mathcal{T}, \Theta_{\mathcal{T}}) = (\mathcal{S}, \Theta_{\mathcal{S}})$ with $(\mathcal{S}, \Theta_{\mathcal{S}})' \subseteq (\mathcal{S}, \Theta_{\mathcal{S}})$. Hence, whilst a naive reading of $(\mathcal{T}, \Theta_{\mathcal{T}})$ might suggest that the event $\Lambda(e)$ happens before $\Lambda(e')$, the swap operation reverses this order. Since $(\mathcal{T}, \Theta_{\mathcal{T}})$ and $(\mathcal{S}, \Theta_{\mathcal{S}})$ are representations of the same model (by Lemma 2), it would thus for example be spurious to posit a temporal or causal order to these events. On the other hand, if (e, e') is not a root-to-leaf path in an adjacency subtree and for all statistically equivalent representations $(\mathcal{S}, \Theta_{\mathcal{S}}) \in [\mathcal{T}, \Theta_{\mathcal{T}}]$ the order of these events is not reversed, then in the model $\mathbb{P}_{(\mathcal{T}, \Theta_{\mathcal{T}})}$ the event $\iota_{\mathcal{T}}^{-1}(\Lambda(e))$ might unambiguously be asserted to happen before $\iota_{\mathcal{T}}^{-1}(\Lambda(e')) \in \sigma(\Omega)$: see [28].

Consider the following application from the longitudinal experiment described in Examples 2 to 5.

4.1. *The statistical equivalence class of a CHDS staged tree.*

A staged tree model of the Christchurch Health and Development Study (CHDS) [11] has been closely analyzed, e.g. in [2, 6], and has been used to describe the

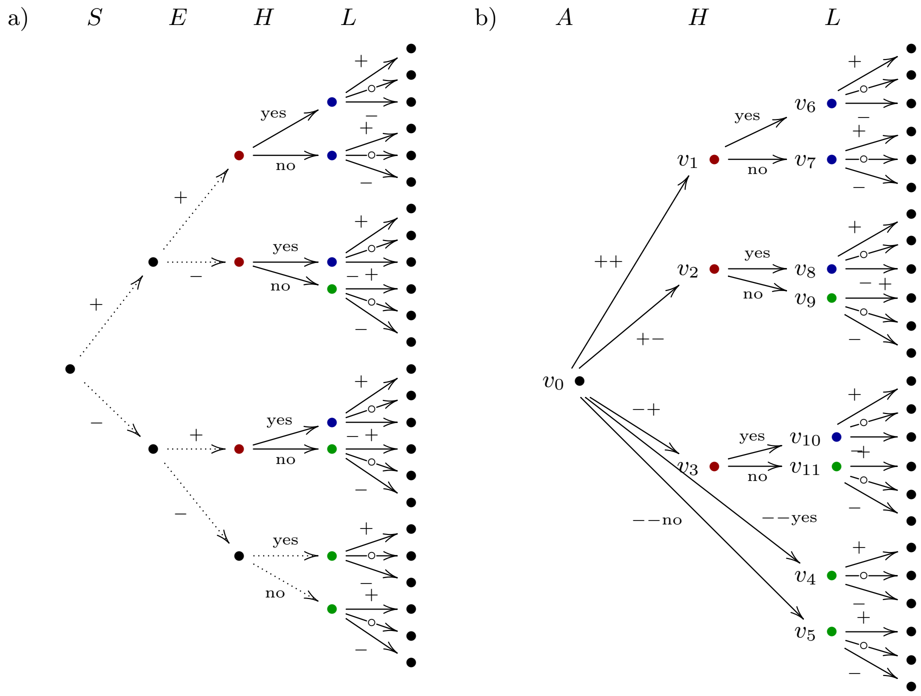

FIGURE 5. a) *The MAP staged tree* $(\mathcal{T}, \Theta_{\mathcal{T}})$ *from* [6], *and* (b) *a statistically equivalent staged tree* $(\mathcal{S}, \Theta_{\mathcal{S}})$. *Here, S, E, H and L are the a priori problem variables and we label edges by the outcomes 'high', 'average' and 'low', indicated as* $+$, $\circ$ *and* $-$, *respectively.*

interplay of the social support S, the economic situation E, hospital admissions H and possible life events L (e.g. divorce) of a group of children in New Zealand over a fixed period of time. Each of these variables is discrete and their simplified state spaces used in [2] are $\mathbb{S} = \mathbb{E} = \{\text{high, low}\}$, $\mathbb{H} = \{\text{yes, no}\}$ and $\mathbb{L} = \{\text{high, average, low}\}$, respectively. In [6], using an MAP search, the authors found the (S, E, H, L)-compatible staged tree $(\mathcal{T}, \Theta_{\mathcal{T}})$ in Fig. 5 (a). We will now apply Theorem 1 to the statistical equivalence class $[\mathcal{T}, \Theta_{\mathcal{T}}]$ in order to enrich our understanding of the model $\mathbb{P}_{(\mathcal{T}, \Theta_{\mathcal{T}})}$.

Note first that there is a saturated subtree $(\mathcal{T}, \Theta_{\mathcal{T}})' \subseteq (\mathcal{T}, \Theta_{\mathcal{T}})$ depicted by dotted lines. By Corollary 1, the associated variables $S \prec E$ are thus ordered in the polynomial equivalence class of $(\mathcal{T}, \Theta_{\mathcal{T}})$. This order cannot be said to have been deduced from the model search. It is therefore helpful for us to transform $(\mathcal{T}, \Theta_{\mathcal{T}})$ into the statistically equivalent staged tree $(\mathcal{S}, \Theta_{\mathcal{S}})$ of Fig. 5 (b), using a resize operator as in Lemma 4 (a) which gives us the necessary flexibility.

The root's edges $e \in E(v_0) \subseteq E$ in this new tree $\mathcal{S} = (V, E)$ can now be assigned a meaning different from the one in $(\mathcal{T}, \Theta_{\mathcal{T}})$. For $e_i = (v_0, v_i)$ and $i = 1, 2, 3$, the embedding $\iota_{\mathcal{S}} : \mathbb{S} \times \mathbb{E} \times \mathbb{H} \times \mathbb{L} \to \Lambda(\mathcal{S})$ interprets e_1, e_2 and e_3 as 'social background or economic status are high' and $e_j = (v_0, v_j)$, $j = 4, 5$, as 'both social background and economic status are low, hospital admission yes or no'. Hence, e_1 denotes a 'child from a wealthy background', e_2 and e_3 as 'from a moderately wealthy background' and e_4 and e_5 as 'from a poor background'. Note from the stages of $(\mathcal{S}, \Theta_{\mathcal{S}})$ that the probabilities of certain numbers of life events are different for wealthy and poor children. Interestingly, [5] names the *access to credit* as a possible monetary measurement of poverty. So being able to borrow from a social network (indicated by S) or having own savings (indicated by E) is a natural indicator of wealth. This gives some external support for moving from $(\mathcal{T}, \Theta_{\mathcal{T}})$ to $(\mathcal{S}, \Theta_{\mathcal{S}})$, suggested from the results of our automated MAP search on the CHDS data. So henceforth let A be the random variable describing access to credit, with state space $\mathbb{A}_1 \cup \mathbb{A}_2 = \iota_{\mathcal{T}}^{-1}(\Lambda(\mathcal{T}'))$ where $\mathbb{A}_1$ is the event that a child is (moderately) wealthy, and $\mathbb{A}_2$ that it is poor in the sense above.

We next analyze the polynomial equivalence class $[\mathcal{S}, \Theta_{\mathcal{S}}]^c$ for $c = c_{\mathcal{S}}$. There are five adjacency subtrees $(\mathcal{S}, \Theta_{\mathcal{S}})_1, \ldots,$ $(\mathcal{S}, \Theta_{\mathcal{S}})_5$ in $(\mathcal{S}, \Theta_{\mathcal{S}})$. These are the ones where v_1, $v_2 \in u_{\text{red}}$, v_1, $v_3 \in u_{\text{red}}$, v_2, $v_3 \in u_{\text{red}}$, v_4, $v_5 \in u_{\text{green}}$ and v_6, $v_7 \in u_{\text{blue}}$ have the same parent and are in the same stage, respectively. Every such subtree depicts a context-specific conditional independence on the problem variables, as stated above and outlined in Section 3.1. There are $2^5 = 32$ possible representations in this polynomial equivalence class. Now, the swaps $\mathfrak{s}_1$, $\mathfrak{s}_2$ and $\mathfrak{s}_3$ which act on $(\mathcal{S}, \Theta_{\mathcal{S}})_1$, $(\mathcal{S}, \Theta_{\mathcal{S}})_2$ and $(\mathcal{S}, \Theta_{\mathcal{S}})_3$, respectively, all change the order of $A \in \mathbb{A}_1$ and H. The swaps $\mathfrak{s}_4$ and $\mathfrak{s}_5$ acting on the subtrees $(\mathcal{S}, \Theta_{\mathcal{S}})_4$ and $(\mathcal{S}, \Theta_{\mathcal{S}})_5$ change the order between $A \in \mathbb{A}_2$ and L. It would thus be spurious to assert a potentially causal or chronological order on these events and random variables.

However, there is no staged tree in the polynomial equivalence class $[\mathcal{S}, \Theta_{\mathcal{S}}]^c$ that would allow for the total order $L \prec H$. This is because no composition of the swaps $\mathfrak{s}_1$, $\ldots$, $\mathfrak{s}_5$ can form a level-swap on $(\mathcal{S}, \Theta_{\mathcal{S}})$. So a model which treats L as an explanation variable of the response variable H as in the study [2] is less supported by the data than one treating H as an explanatory variable of L as in [6]. Of course this deduction needs the caveat that there exists a reasonably high scoring staged tree model which *does* embed this reversal. So evidence for the chosen order is quite weak. However, it is nevertheless formally *suggested* in the unambiguous way we discuss above. Note that no deductions about an ordering of variables were possible within the original BN representation of the data because the MAP model turns out to be decomposable. This demonstrates that the extra structure of the staged tree enables us to draw out new potential causal hypotheses that could not be discovered when using more conventional graphical methods.

The statistical equivalence class $[\mathcal{T}, \Theta_{\mathcal{T}}]$ is very large. It contains nearly one thousand elements. In particular, through examining how the stratification can be maintained using certain resize operators, we can make more specific assertions about the interplay between hospital admissions and life events, conditioning on whether or not a child is poor. Similarly, if we first swap and then resize saturated subtrees in $(\mathcal{T}, \Theta_{\mathcal{T}})$, we can obtain statements on new variables (S, H) and E. The statistical equivalence class of the MAP tree found in [2] and analyzed in Examples 2 to 5, where during the search the order of variables was restricted to $S \prec E \prec L \prec H$, is similarly rich: we count more than five hundred elements in the polynomial equivalence class of one particular staged tree representation alone.

Thus, knowing the graphical structure of all elements in the statistical equivalence class of a staged tree enables us to analyze all representations of the same model, so all different equivalent ways of explaining the data. Furthermore, although we do not give details here, knowing that so many different graphs actually all represent the same model can significantly speed up the search across these representations of a set of problem variables: performing model search on the equivalence classes rather than on the staged trees themselves.

5. Discussion.

In this paper we have been able to show that a characterization of staged trees in terms of their interpolating polynomial provides an elegant way to fully analyze equivalence classes of these models.

Throughout, we have been pointing to putative causal interpretations. It has been argued by other authors, e.g. [28], that in fact causal hypotheses are most easily drawn from and analyzed in tree graphs. In a forthcoming publication we will use the results above as the basis for developing a more detailed analysis of tree-based causality in comparison to analogous concepts developed for BN models [23, 32].

Similarly, the theory developed in this paper naturally extends into the domain of algebraic and differential geometry, just as with BN models [9]. In a different publication, algebra will further help us to define a fast algorithm to construct graphical model representations from a given interpolating polynomial.

We believe that in the future this will provide very promising grounds of research that will help guide statistical inference in general and causal inference in particular.

APPENDIX A: PROOFS OF RESULTS

Proof of Lemma 2. Let $(\mathcal{T},\Theta_{\mathcal{T}})$ and $(\mathcal{S},\Theta_{\mathcal{S}})$ be polynomially equivalent. Set $g_\omega = \mathbb{1}_\omega$ for any $\omega \in \Omega$. By Definition 4, the network polynomials $c_{g_\omega,\mathcal{S}} = c_{g_\omega,\mathcal{T}}$ are equal and hence

$$c_{\mathbb{1}_\omega,\mathcal{S}} = \pi_{\theta,\mathcal{S}}(\iota_{\mathcal{S}}(\omega)) = \pi_{\theta,\mathcal{T}}(\iota_{\mathcal{T}}(\omega)) = c_{\mathbb{1}_\omega,\mathcal{T}} \tag{A.1}$$

for all $\omega \in \Omega$. Hence, the terms of $c_{g_\omega,\mathcal{S}}$ and $c_{g_\omega,\mathcal{T}}$ coincide termwise. Thus, by the evaluation homomorphism, the atomic probabilities also coincide pairwise. So the functions $\pi_{\theta,\mathcal{S}} = \pi_{\theta,\mathcal{T}}$ are equal. Finally, by definition, $(\mathcal{T},\Theta_{\mathcal{T}})$ and $(\mathcal{S},\Theta_{\mathcal{S}})$ are statistically equivalent. □

Proof of Proposition 1. For saturated models, a parametrization in terms of atomic probabilities is tree compatible. So in this case Example 1 yields the claim. Assume now $\mathbb{P}_\Psi$ is not saturated and Ψ is multilinear.

Sufficiency of the first claim is straight forward. Indeed, by Definition 5 and the nested factorization in (3.10), the interpolating polynomial of a probability tree model is tree compatible.

For necessity assume now the interpolating polynomial $c = c_{1,\Psi} \in \mathbb{R}_d[\Theta]$ of $\mathbb{P}_\Psi$ is tree compatible and given by the factorization $s(c(\theta))$ in (3.11).

Let $\mathcal{T} = (V,E)$ be a directed tree whose florets are labelled by the subsums of (3.11), $\mathcal{F}_j = (v_j, \{e \mid \theta(e) = \theta_j \in A_j(\theta_{j-1})\})$, $j \in [k]$, and partially ordered by reversing the steps that gave us (3.10) above. Then we can define a map $\mathfrak{c} : s(c(\theta)) \mapsto (\mathcal{T},\Theta_{\mathcal{T}})$. This inductively labels leaf-floret edges in $\mathcal{T}$ by the innermost factors $A_k(\theta_{k-1})$ of $s(c(\theta))$ and the roots' edges by the outermost factors A_1. Since by definition every set $A_j(\theta_{j-1})$ has at least two elements, it follows that there are at least two edges in every floret. So the tree-compatible factorizations in (3.10) and (3.11) are equal and $(\mathcal{T},\Theta_{\mathcal{T}})$ is a probability tree with $c = c_{\mathcal{T}}$. Let $(\Omega, \sigma(\Omega), P_\theta)$ denote the space underlying $\mathbb{P}_\Psi$. If Ψ is multilinear then P_θ and thus $\pi_{\theta,\mathcal{T}}$ are (algebraically) injective functions. Using the identification of root-to-leaf paths $\lambda \in \Lambda(\mathcal{T})$ with $\omega \in \Omega$ such that $P_\theta(\omega) = \pi_{\theta,\mathcal{T}}(\iota_{\mathcal{T}}(\omega))$, it can be seen that $(\mathcal{T},\Theta_{\mathcal{T}})$ is indeed a representation of the model at hand and $\Psi = \Psi_{\mathcal{T}}$ as well as $\mathbb{P}_\Psi = \mathbb{P}_{(\mathcal{T},\Theta_{\mathcal{T}})}$.

The map $\mathfrak{c}$ is injective because, by (3.10), two probability trees with differently labelled graphs give rise to a different order of summation. By the construction above, clearly $\mathrm{im}(\mathfrak{c}) \subseteq [\mathcal{T},\Theta_{\mathcal{T}}]^c$. Moreover, $\mathrm{im}(\mathfrak{c}) = [\mathcal{T},\Theta_{\mathcal{T}}]^c$ because every representation $(\mathcal{T},\Theta_{\mathcal{T}}) \in [\mathcal{T},\Theta_{\mathcal{T}}]^c$ has c as its interpolating polynomial. Thus, $\mathfrak{c} : \{s(c(\theta)) \mid s\} \to [\mathcal{T},\Theta_{\mathcal{T}}]^c$ is bijective. □

Proof of Proposition 2. First let $(\mathcal{T},\Theta_{\mathcal{T}}) = \mathfrak{c}(s_1(c(\theta)))$ and $(\mathcal{S},\Theta_{\mathcal{S}}) = \mathfrak{c}(s_2(c(\theta)))$ be polynomially equivalent staged trees with common interpolating polynomial c and corresponding factorizations s_1 and s_2. $\mathfrak{c}$ denotes the map from Proposition 1. Clearly, one factorization $s_1(c(\theta))$ is transformed into the other $s_2(c(\theta))$ by applying the distributive law of $+$ and $\cdot$ a finite number of times. Hence, we can define a map $\tilde{s} : s_1(c(\theta)) \mapsto s_2(c(\theta))$ performing these calculations on the subsums of c as in (3.13). Therefore,

$$\mathfrak{s} :\ (\mathcal{T}, \Theta_{\mathcal{T}}) \overset{\mathfrak{c}^{-1}}{\mapsto} s_1(c(\theta)) \overset{\tilde{s}}{\mapsto} s_2(c(\theta)) \overset{\mathfrak{c}^{-1}}{\mapsto} (\mathcal{S}, \Theta_{\mathcal{S}}) \tag{A.2}$$

is a map which performs a finite number of swaps on the to $\tilde{s}$ corresponding adjacency subtrees and thus transforms $(\mathcal{T}, \Theta_{\mathcal{T}})$ into $(\mathcal{S}, \Theta_{\mathcal{S}})$.

Conversely, let $(\mathcal{T}, \Theta_{\mathcal{T}})$ be any staged tree. By Definition 6, $\mathfrak{s}_1(\mathcal{T}, \Theta_{\mathcal{T}})$ is polynomially equivalent to $(\mathcal{T}, \Theta_{\mathcal{T}})$. It trivially follows that a finite number of swaps yield a tree $(\mathcal{S}, \Theta_{\mathcal{S}}) = \mathfrak{s}_l \circ \mathfrak{s}_{l-1} \circ \ldots \circ \mathfrak{s}_1(\mathcal{T}, \Theta_{\mathcal{T}})$ which is polynomially equivalent to $(\mathcal{T}, \Theta_{\mathcal{T}})$. □

Proof of Lemma 3. As also noticed by [8, 15],

$$c_{\mathbb{1}_A, \mathcal{T}}(\theta) = \sum_{\lambda \in \Lambda(\mathcal{T})} \mathbb{1}_A(\lambda) \pi_{\theta, \mathcal{T}}(\lambda) = \pi_{\theta, \mathcal{T}}(A) = P_\theta(\iota_{\mathcal{T}}^{-1}(A)) \tag{A.3}$$

for any $A \in \sigma(\mathcal{T})$. Thus, as a function $c_{\mathbb{1}_\cdot, \mathcal{T}}$ coincides with the measure P_θ on the probability space $(\Omega, \sigma(\Omega), P_\theta)$ underlying $\mathbb{P}_{(\mathcal{T}, \Theta_{\mathcal{T}})}$, independent of the embedding $\iota_{\mathcal{T}}$ used to identify that space with $(\Lambda(\mathcal{T}), \sigma(\mathcal{T}), \pi_{\theta, \mathcal{T}})$. By definition, for every two statistically equivalent probability trees this measure is the same. □

Proof of Lemma 4. a) Since the image $\mathfrak{r}(\mathcal{T}, \Theta_{\mathcal{T}}) = (\mathcal{S}, \Theta_{\mathcal{S}})$ of a staged tree is a probability tree and since by assumption the non-trivial stage sets of image and preimage coincide, $\tilde{U}_{\mathcal{T}} = \tilde{U}_{\mathcal{S}}$, clearly also $(\mathcal{S}, \Theta_{\mathcal{S}}) \in [\mathcal{T}, \Theta_{\mathcal{T}}]$ is a staged tree.

b) The assumptions in this case imply that the stage-structure of the naively resized subgraphs $(\mathcal{T}, \Theta_{\mathcal{T}})' \subseteq (\mathcal{T}, \Theta_{\mathcal{T}})$ is self-contained in the sense that if we can show that $(\mathcal{S}, \Theta_{\mathcal{S}})' = \mathfrak{r}(\mathcal{T}, \Theta_{\mathcal{T}})'$ is a staged tree, then there are no extra constraints within $(\mathcal{T}, \Theta_{\mathcal{T}})$ or $(\mathcal{S}, \Theta_{\mathcal{S}}) = \mathfrak{r}(\mathcal{T}, \Theta_{\mathcal{T}})$ which could create non-linear structure. Now, because all subgraphs $(\mathcal{T}, \Theta_{\mathcal{T}})'$, $(\mathcal{T}, \Theta_{\mathcal{T}})'' \subseteq (\mathcal{T}, \Theta_{\mathcal{T}})$ that $\mathfrak{r}$ acts on are polynomially equivalent, we find in $\mathfrak{r}(\mathcal{T}, \Theta_{\mathcal{T}})'$ and $\mathfrak{r}(\mathcal{T}, \Theta_{\mathcal{T}})''$ that the atomic probabilities $\pi_{\theta, \mathcal{T}'}(\lambda') = \pi_{\theta, \mathcal{T}''}(\lambda'')$ coincide for subpaths λ', λ'' which have the same atomic monomial in $(\mathcal{T}, \Theta_{\mathcal{T}})$. Thus, the image $(\mathcal{S}, \Theta_{\mathcal{S}}) = \mathfrak{r}(\mathcal{T}, \Theta_{\mathcal{T}})$ is a staged tree where the stages are given by these identified (formerly atomic now) primitive probabilities. □

Proof of Theorem 1. First let $(\mathcal{T}, \Theta_{\mathcal{T}})$, $(\mathcal{S}, \Theta_{\mathcal{S}}) \in [\mathcal{T}, \Theta_{\mathcal{T}}]$ be statistically equivalent staged trees. By (2.9), for all identified root-to-leaf paths

$\lambda' = \iota_{\mathcal{S}}(\iota_{\mathcal{T}}(\lambda))$ their atomic probabilities $\pi_{\theta,\mathcal{T}}(\lambda) = P_\theta(\iota_{\mathcal{T}}^{-1}(\lambda)) = \pi_{\theta',\mathcal{S}}(\lambda')$ are equal. Here, P_θ denotes the underlying measure on Ω that $(\mathcal{T}, \Theta_{\mathcal{T}})$ and $(\mathcal{S}, \Theta_{\mathcal{S}})$ have in common. If the above equality holds in a formal sense for every $\lambda \in \Lambda(\mathcal{T})$ then $(\mathcal{T}, \Theta_{\mathcal{T}})$ and $(\mathcal{S}, \Theta_{\mathcal{S}})$ are polynomially equivalent. In this case, Lemma 2 states that a map exists between the two staged trees which is a composition of swaps, and thus proves the claim. If this is not the case, we denote by $\Lambda \subseteq \Lambda(\mathcal{T})$ the set of root-to-leaf paths in $\mathcal{T}$ whose atomic monomials do not coincide formally with the corresponding atomic monomials in $\mathcal{S}$. Let $(\mathcal{T}, \Theta_{\mathcal{T}})' \subseteq (\mathcal{T}, \Theta_{\mathcal{T}})$ denote a subtree of $\mathcal{T}$ for which $\Lambda \subseteq \Lambda(\mathcal{T}')$, and define analogously the corresponding $(\mathcal{S}, \Theta_{\mathcal{S}})' \subseteq (\mathcal{S}, \Theta_{\mathcal{S}})$. These are the subtrees which are not polynomially equivalent, and thus have different parametrizations. We define two resize operators, $\mathfrak{r}_{\mathcal{T}} : (\mathcal{T}, \Theta_{\mathcal{T}})' \mapsto (\mathcal{F}, \{\theta_{\mathcal{F}}\})$ and $\mathfrak{r}_{\mathcal{S}} : (\mathcal{S}, \Theta_{\mathcal{S}})' \mapsto (\mathcal{F}, \{\theta_{\mathcal{F}}\})$ which map those subtrees to the same floret. By Lemma 5, $(\mathcal{S}, \Theta_{\mathcal{S}})'$, $(\mathcal{T}, \Theta_{\mathcal{T}})'$ and $(\mathcal{F}, \{\theta_{\mathcal{F}}\})$ are statistically equivalent. Thus, there is a composition of resizes $\mathfrak{r} = \mathfrak{r}_{\mathcal{S}}^{-1} \circ \mathfrak{r}_{\mathcal{T}} : (\mathcal{T}, \Theta_{\mathcal{T}}) \mapsto (\mathcal{S}, \Theta_{\mathcal{S}})$ between the statistically equivalent staged trees.

Now let $\mathfrak{m}$ be a transformation given by swaps and resizes between two staged trees $(\mathcal{T}, \Theta_{\mathcal{T}})$, $(\mathcal{S}, \Theta_{\mathcal{S}})$. If $\mathfrak{m}$ is a composition of swaps, then Proposition 1 ensures polynomial equivalence, and thus statistical equivalence by Lemma 2. If $\mathfrak{m}$ is a composition of resizes, then Lemma 5 yields statistical equivalence. Clearly, also for the composition of both of these operators holds that $(\mathcal{T}, \Theta_{\mathcal{T}})$ and $\mathfrak{m}(\mathcal{T}, \Theta_{\mathcal{T}}) = (\mathcal{S}, \Theta_{\mathcal{S}})$ are statistically equivalent. The claim follows. □

ACKNOWLEDGEMENTS

Christiane Görgen was supported by the EPSRC grant EP/L505110/1.

REFERENCES

[1] Andersson, S. A., Madigan, D. and Perlman, M. D. (1997). A Characterization of Markov Equivalence Classes for Acyclic Digraphs. *Annals of Statistics* **25** 505-541.

[2] Barclay, L. M., Hutton, J. L. and Smith, J. Q. (2013). Refining a Bayesian Network using a Chain Event Graph. *Internat. J. Approximate Reasoning* **54** 1300-1309.

[3] Boutilier, C., Friedman, N., Goldszmidt, M. and Koller, D. (1996). Context-specific Independence in Bayesian Networks. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence. UAI'96* 115-123. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[4] Collazo, R. A. and Smith, J. Q. (2015). A new family of Non-Local Priors for Chain Event Graph model selection. *CRiSM* **05-02**.

[5] Coudouel, A., Hentschel, J. S. and Wodon, Q. T. (2002). *Poverty Measurement and Analysis* In *A sourcebook for Poverty Reduction Strategies: Core techniques and cross-cutting issues* 27-74. The World Bank.

[6] Cowell, R. G. and Smith, J. Q. (2014). Causal discovery through MAP selection of stratified Chain Event Graphs. *Electronic Journal of Statistics* **8** 965-997.

[7] Cox, D. A., Little, J. B. and O'Shea, D. (1997). *Ideals, Varieties and Algorithms*, 2nd ed. Springer.

[8] Darwiche, A. (2003). A differential approach to inference in Bayesian networks. *Journal of ACM* **50**.

[9] Drton, M., Sturmfels, B. and Sullivant, S. (2009). *Lectures on Algebraic Statistics. Oberwolfach Seminars; Vol. 39.* Birkhäuser Verlag, Basel.

[10] Drton, M. and Sullivant, S. (2007). Algebraic Statistical Models. *Statistica Sinica* **17** 1273-1297.

[11] Fergusson, D. M., Horwood, L. J. and Shannon, F. T. (1986). Social and family factors in childhood hospital admission. *Journal of Epidemiology and Community Health* **40** 50-58.

[12] Freeman, G. and Smith, J. Q. (2011). Bayesian MAP model selection of Chain Event Graphs. *J. Multivariate Anal.* **102**.

[13] Geiger, D., Meek, C. and Sturmfels, B. (2006). On the toric algebra of graphical models. *Annals of Statistics* **34** 1463-1492.

[14] Gibilisco, P., Riccomagno, E., Rogantin, M. P. and Wynn, H. P., eds. (2010). *Algebraic and Geometric Methods in Statistics*, 1st ed. Cambridge University Press.

[15] Görgen, C., Leonelli, M. and Smith, J. Q. (2015). A Differential Approach for Staged Trees. In *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty.*

[16] Heckerman, D. (1998). *A Tutorial on Learning with Bayesian Networks* In *Learning in Graphical Models* 301-354. The MIT Press.

[17] Jensen, F. V. and Jensen, F. (1994). Optimal Junction Trees. In *Proceedings of the 10th Conference on Uncertainty in Artifical Intelligence* (M. Kaufmann, ed.) **10**.

[18] Lauritzen, S. L. (1996). *Graphical Models. Oxford Statistical Science Series.* Oxford University Press, New York.

[19] Lauritzen, S. L. and Richardson, T. S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society, Series B* **64** 321-348.

[20] Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems. *Journal of the Royal Statistical Society, Series B* **50** 157-224.

[21] Lauritzen, S. L., Dawid, P. A., Larsen, B. N. and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks* **20** 491-505.

[22] Leonelli, M., Görgen, C. and Smith, J. Q. (2016). Sensitivity Analysis for Staged Trees. *In preparation.*

[23] Pearl, J. (2000). *Causality*, 2nd ed. Cambridge University Press.

[24] Pistone, G., Riccomagno, E. and Wynn, H. P. (2001a). *Algebraic Statistics. Monographs on Statistics and Applied Probability; Vol. 89.* Chapman & Hall/CRC, Boca Raton, FL Computational commutative algebra in statistics.

[25] Pistone, G., Riccomagno, E. and Wynn, H. P. (2001b). Gröbner bases and factorisation in discrete probability and Bayes. *Statistics and Computing* **11** 37-46.

[26] Salmerón, A., Cano, A. and Moral, S. (2000). Importance sampling in Bayesian networks using probability trees. *Computational Statistics and Data Analysis* **34** 387-413.

[27] Schachter, R. D. (1988). Probabilistic Inference and Influence Diagrams. *Operations Research* **36** 589-605.

[28] Shafer, G. (1996). *The Art of causal Conjecture.* MIT Press, Cambridge.

[29] Smith, J. Q. (2010). *Bayesian Decision Analysis, Principles and Practice.* Cambridge

University Press.

[30] Smith, J. Q. and Anderson, P. E. (2008). Conditional independence and Chain Event Graphs. *Artifical Intelligence* **172** 42-68.

[31] Studený, M. (2005). *Probabilistic Conditional Independence Structures. Information Science and Statistics.* Springer.

[32] Thwaites, P. A. (2013). Causal identifiability via Chain Event Graphs. *Artifical Intelligence* **195** 291-315.

[33] Thwaites, P. A., Smith, J. Q. and Cowell, R. G. (2008). Propagation using Chain Event Graphs. In *Proceedings of the 24th Conference on Uncertainty in Artifical Intelligence* 546-553.

[34] Thwaites, P. A., Smith, J. Q. and Riccomagno, E. (2010). Causal Analysis with Chain Event Graphs. *Artificial Intelligence* **174** 889-909.

[35] Thwaites, P. A. and Smith, J. Q. (2015). A Separation Theorem for Chain Event Graphs. *arXiv:1501.05215.*

[36] Ye, K. Q. (2003). Indicator Function and Its Application in Two-Level Factorial Designs. *Annals of Statistics* **31** 984-994.

Department of Statistics
University of Warwick
Coventry CV4 7AL
United Kingdom
E-mail: c.gorgen@warwick.ac.uk
j.q.smith@warwick.ac.uk