JMLR: Workshop and Conference Proceedings vol 52, 207-215, 2016 PGM 2016

A Differential Approach to Causality in Staged Trees

Christiane Görgen C.GORGEN@WARWICK.AC.UK
Jim Q. Smith J.Q.SMITH@WARWICK.AC.UK
Department of Statistics
University of Warwick
Coventry CV4 7AL, United Kingdom

Abstract

In this paper, we apply a recently developed differential approach to inference in staged tree models to causal inference. Staged trees generalise modelling techniques established for Bayesian networks (BN). They have the advantage that they can depict highly nuanced structure impossible to express in a BN and also enable us to perform causal manipulations associated with very general types of interventions on the system. Conveniently, what we call the interpolating polynomial of a staged tree has been found to be an analogue to the essential graph of a BN. By analysing this polynomial in a differential framework, we find that interventions on the model can be expressed as a very simple operation. We can therefore clearly state causal hypotheses which are invariant for all staged trees representing the same causal model. The technology we develop here, illustrated through a simple example, enables us to search for a variety of complex manipulations in large systems accurately and efficiently.

Keywords: Causality; causal manipulation; differential approach; generating functions; probability trees; staged trees.

1. Introduction

A *staged tree* is a probability tree equipped with graphical context-specific conditional independence information which represents a discrete parametric statistical model (Smith and Anderson, 2008). This at a first glance rather complex graph enables the statistician to accurately represent asymmetric domains through a simple and harmonious set of semantics. In contrast, BN models often unnecessarily need to embed an underlying state space within a product structure on a given set of random variables and add non-graphical information to capture constraints on the parameter space. Staged trees are thus not only generalisations of discrete BNs, they also do not rely on a set of problem variables specified a priori. This makes them especially useful in modelling problems where the focus is on how events might unfold rather than how random variables might be related (Barclay et al., 2013). Fast inference techniques for these models are now available (Freeman and Smith, 2011; Collazo and Smith, 2015). In addition, staged trees can compactly and graphically represent a vast variety of causal manipulations not amenable to a composite representation of atomic manipulations associated with the corresponding BN. Causal interventions are especially amenable to the staged tree framework where they more naturally concern events rather than random variables.

The usefulness of probability trees as graphical representations of causal conjectures was first properly articulated by Shafer (1996). Subsequently, the mathematical apparatus for expressing a *causal manipulation*—or control—of staged trees was then successfully introduced by Riccomagno and Smith (2005); Thwaites et al. (2010); Thwaites (2013). Here, the definitions of this type of manipulation were chosen to be analogous to Pearl's do-operator in BNs (Pearl, 2000). Through

this new semantic it was possible to show that Pearl's implicit embedding could be extended to non product space structures in a straight forward way.

One of our central observations which lead to this work is that the algebraic characterisation of staged trees in Görgen and Smith (2015) can be combined with the results of Riccomagno and Smith (2009): the manipulation operation on staged trees is in fact an algebraic operation. It has already been shown that methods proposed by Darwiche (2003) that used techniques of differential calculus to analyse the structure of BNs could also be applied to staged trees (Görgen et al., 2015). Here, we show that the same framework applies to causal manipulations. These can be interpreted as a new local differentiation operation. This operation is a simple but powerful tool which makes these calculations amenable to the use of computer algebra and fast investigation of many candidate complex interventions. Furthermore, because of a certain factorisation of the characterising polynomial, the effects of causal manipulation and the strength of causal hypotheses can be efficiently and speedily computed.

The polynomial on which we perform this operation indexes the equivalence class of all staged trees representing the same statistical model. Operations on this polynomial are thus analogous to interventions on the essential graph of a BN. So through this new framework, results developed above are invariant to a graphical representation we choose for a statistical model and can be easily applied to parametric statistical models which are even more general than staged trees (Leonelli et al., 2015). These methods provide us with exciting new possibilities for graphically based causal representations which apply outside the class of problems where an observed system can be faithfully described using a single BN.

2. Causal Manipulation in Staged Tree Models

In Görgen and Smith (2015), the pair $(\mathcal{T}, \Theta_\mathcal{T})$ is called a *staged tree* if the graph $\mathcal{T} = (V, E)$ is an event tree together with a set $\Theta_\mathcal{T} = \{\boldsymbol{\theta}_v \mid v \in V\}$ of vectors $\boldsymbol{\theta}_v = (\theta(e) \mid e \in E(v))$ of parameters for each vertex $v \in V$ with emanating edges $E(v) = \{e \in E \mid e = (v, \cdot)\}$. These parameter vectors always lie in open probability simplicies: their entries sum to unity, $\sum_{e \in E(v)} \theta(e) = 1$ for all $v \in V$, and are positive, $\theta(e) \in (0, 1)$ for all $e \in E$. Whenever $\boldsymbol{\theta}_v = \boldsymbol{\theta}_w$ then v and w are said to be in the same non-trivial *stage*. This type of parameter identification can be thought of as setting conditional probabilities equal in a (context-specific) BN. Every root-to-leaf path $\lambda \in \Lambda(\mathcal{T})$ in such a *staged tree* is a sequence of edges $e \in E(\lambda)$ and represents a possible unfolding of events in the context represented by $(\mathcal{T}, \Theta_\mathcal{T})$. The probability of every such atom $\lambda \in \Lambda(\mathcal{T})$ can be calculated as the product of its edge-labels, so is a monomial

$$\pi_{\boldsymbol{\theta},\mathcal{T}}(\lambda) = \prod_{e \in E(\lambda)} \theta(e). \tag{1}$$

For simplicity we focus in this paper on staged trees that do not have repetitive stage structure along root-to-leaf paths, so those whose atomic probabilities are square-free monomials.

Note that every staged tree $(\mathcal{T}, \Theta_\mathcal{T})$ represents a discrete parametric model whose elements are probability mass functions $\pi_{\boldsymbol{\theta},\mathcal{T}}$ whose parametrisation can be read from the graph together with its labels.

Now let $(\mathcal{T}, \Theta_\mathcal{T})$ be a given staged tree representing a population before any intervention takes place. Following Pearl (2000), we call this the *idle* system. Then Riccomagno and Smith (2005);

Thwaites (2013) define the *manipulated* staged tree to be a subtree of a the idle tree with inherited edge-labels. Most often intervention is performed on *vertex-centred* events $\Lambda(v) = \{\lambda \in \Lambda(\mathcal{T}) \mid \text{exists } e = (v, \cdot) \in E(\lambda)\}$ given by all root-to-leaf paths going through a fixed vertex $v \in V$. So these events contain all atoms for which the interpretation of 'reaching v' is true within the modelling context. The resulting manipulated subtree $(\mathcal{T}(v), \Theta_{\mathcal{T}(v)})$ is then rooted at v in $\mathcal{T}$ and its root-to-leaf paths are exactly the v-to-leaf paths in $\Lambda(\mathcal{T})$ with their respective labels, so $\Theta_{\mathcal{T}(v)} \subseteq \Theta_{\mathcal{T}}$. This manipulated tree now represents a setting where a unit from the idle system is controlled so that it is forced to go through the vertex v. The causal hypothesis then simply asserts that the subtree of the staged tree describing the future development of that unit at that vertex is the same as it would be were that unit to have arrived in that situation naturally, but that the earlier development of the unit before it reaches the vertex will remain unchanged by the control.

In terms of the probability mass function, this intervention corresponds to replacing the probability of an event in the idle system with that probability in the manipulated subtree:

$$\pi_{\boldsymbol{\theta},\mathcal{T}}(\lambda \mid\mid \Lambda(v)) = \pi_{\boldsymbol{\theta},\mathcal{T}(v)}(\lambda) \tag{2}$$

for $v \in V$ and $\mathcal{T}(v) \subseteq \mathcal{T} = (V, E)$. See Thwaites (2013) for a more detailed discussion.

Staged trees are very general models. In particular, every acyclic digraph representation of a discrete and finite BN has an equivalent staged tree representation (Smith and Anderson, 2008). But staged trees can express many different and useful manipulations. For example in Pearl (1995), atomic causal manipulations on BN models are always a composition of controls of the form $\mathrm{do}(X = x)$, where this *do*-operator expresses the control that a random variable X is externally forced to take a certain value x. There are many circumstances when we might need to consider the impact of different classes of interventions: for instance, we might want to assign certain treatments $X = x$ exclusively to patients with a particular history $Y = y$. This would require a conditional or context-specific manipulation $\mathrm{do}(X_{Y=y} = x)$ where $X_{Y=y}$ is the conditional random variable $X|Y = y$. In a staged tree representation of such a BN model, Pearl's intervention $\mathrm{do}(X = x)$ would correspond to a simultaneous intervention on all vertices associated with X, typically lying at the same distance from the root in a *stratified* staged tree (Thwaites et al., 2010; Cowell and Smith, 2014). So this would force $X = x$ independent of the context. In addition, these interventions are usually performed on a causal graph, possibly inferred using an appropriate causal discovery algorithm, rather than on a causal equivalence class of representations where both $X \to Y$ and $Y \leftarrow X$ might be valid directionalities.

In contrast, a single vertex intervention as proposed above is far more flexible and, choosing one particular situation, can manipulate $X = x$ only when $Y = y$. Although contingent manipulations have subsequently been studied in BNs, causal manipulation can be expressed much more simply within staged tree representations. Furthermore, within this framework the semantics defined also extend seamlessly to models that cannot naturally be embedded into a product space (see our running example). In particular, staged trees do not rely on a set of a priori problem variables, so manipulation can instead be analysed in terms of any events of interest. We show in this paper that vertex manipulations can be translated into a local differential operation which is analogous to an intervention on the essential graph of a BN.

Thus, consider an example motivated by Thwaites (2013). First year students at a university either live on campus or in one of two cities nearby. Landlords in both cities can be either friendly or grumpy. If they are grumpy, then students might consider moving house, and if they do move

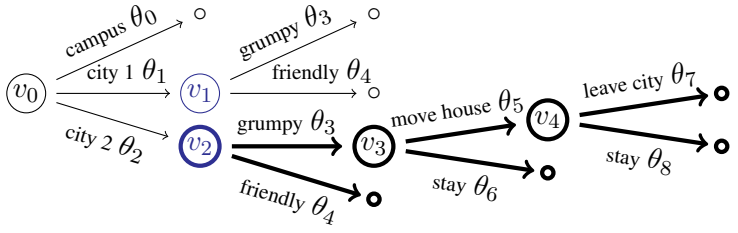

Figure 1: A staged tree $(\mathcal{T}, \Theta_{\mathcal{T}})$ depicting an idle system. Its thick subtree $(\mathcal{T}(v_2), \Theta_{\mathcal{T}(v_2)})$ depicts the unfolding when the event $\Lambda(v_2)$ is enforced. See page 209 for an analysis.

then they might also consider leaving the city. We shall assume that the attitude of landlords is the same in both cities, such that the probability of renting with a friendly landlord does not depend on the location. We also assume that we are only interested in the flow of students in one of the two cities. Rather than transforming this process into a BN which would require us to first define a set of problem variables with a product state space with many redundant entries in the conditional probability tables, we can better analyse the asymmetric constraints in a staged tree as in fig. 1. Note here that vertices v_1 and v_2 are in the same stage because their attached transition probabilities are equal by assumption.

Suppose we are now interested in what would happen were students *forced* to move to the city of interest, for instance by a university policy. In that case, the edge (v_0, v_2) would be the only possible unfolding from the root and thus the thick-depicted subtree $(\mathcal{T}(v_2), \Theta_{\mathcal{T}(v_2)})$ in fig. 1 represents the manipulated system as part of the idle system. If our setting was part of a bigger system, this vertex-centred manipulation would be especially sensible because we would not want to force students to live in our city of interest if they went to university in a different part of the country (as the corresponding atomic manipulation in a BN model would imply).

Note that the atomic probabilities depicted in $(\mathcal{T}, \Theta_{\mathcal{T}})$ are given by the monomials θ_0, $\theta_1\theta_3$, $\theta_1\theta_4$, $\theta_2\theta_3\theta_5\theta_7$, $\theta_2\theta_3\theta_5\theta_8$, $\theta_2\theta_3\theta_6$ and $\theta_2\theta_4$ as in eq. (1). So by eq. (2), the projection of $(\mathcal{T}, \Theta_{\mathcal{T}})$ onto $(\mathcal{T}(v_2), \Theta_{\mathcal{T}(v_2)})$ then yields new monomials $\theta_3\theta_5\theta_7$, $\theta_3\theta_5\theta_8$, $\theta_3\theta_6$ and θ_4.

3. Causal Manipulation in a Differential Framework

We will now embed the formalism introduced above into the differential framework from Darwiche (2003). In order to do this, we combine the recent results from Görgen and Smith (2015) with the idea of Riccomagno and Smith (2009); Görgen et al. (2015) to express causal manipulation in terms of projections and operations on a polynomial.

We will call a symbolic sum of the atomic probabilities of a staged tree the *interpolating polynomial* of an event $A \subseteq \Lambda(\mathcal{T})$ in the model represented by $(\mathcal{T}, \Theta_{\mathcal{T}})$, so

$$c_{\mathcal{T},A}(\boldsymbol{\theta}) = \sum_{\lambda \in \Lambda(\mathcal{T})} \prod_{e \in E(\lambda)} \mathbb{1}_{\Lambda(e) \cap A}(\lambda)\theta(e) \tag{3}$$

where $\boldsymbol{\theta}$ is a vector of all edge-labels and $\Lambda(e) = \{\lambda \in \Lambda(\mathcal{T}) \mid e \in E(\lambda)\}$ are *edge-centred* events (Görgen et al., 2015). If $A = \Lambda(\mathcal{T})$, we also write $c_{\mathcal{T}} = c_{\mathcal{T},\Lambda(\mathcal{T})}$ for the interpolating polynomial

of the whole space. Note that eq. (3) is a formal polynomial and that it does not a priori require us to plug in the summing-to-unity conditions of the underlying probability simplex. In Görgen and Smith (2015) it has been shown that this embedding is formally unambiguous.

Interpolating polynomials in staged trees were motivated by the seminal work of Chan and Darwiche (2002); Darwiche (2003). There, analogous objects were defined for BNs and successfully used for performing sensitivity analysis and for calculating conditional and marginal probabilities in these models. This was because the authors found that the probability of an event in the model is simply a polynomial, $\pi_{\boldsymbol{\theta},\mathcal{T}}(A) = c_{\mathcal{T},A}(\boldsymbol{\theta})$. This simple but powerful observation can in fact be applied to a much more general scenario. Here, we will show how it extends to the study of the effects of causal manipulations within the class of staged trees.

By Görgen and Smith (2015), the interpolating polynomial of a staged tree is a central object in the study of statistical equivalence classes. In particular, the class of all staged trees representing the same model can be traversed using two algebra-based operators, called the *swap* and the *resize*. The former is an analogue to arc reversals in BNs and the latter is analogous to using a clique-parametrisation in decomposable BNs. These operations require an interpolating polynomial to be written in terms of a nested factorisation

$$c_{\mathcal{T}}(\boldsymbol{\theta}) = \sum_{(v_0,v_1)\in E(v_0)} \theta(v_0,v_1) \left(\sum_{(v_1,v_2)\in E(v_1)} \theta(v_1,v_2) \left(\cdots \left(\sum_{(v_{k-1},v_k)\in E(v_{k-1})} \theta(v_{k-1},v_k) \right) \right) \right) \tag{4}$$

where $\theta(v,v') = \theta(e)$ are the edge labels of $e = (v,v') \in E$, $v_0 \in V$ is the root and $k \in \mathbb{N}$ is the number of edges of the longest root-to-leaf path in $\mathcal{T} = (V,E)$. Centrally, eq. (4) is in one-to-one correspondence with the labelled graph $(\mathcal{T},\Theta_{\mathcal{T}})$, so one can be deduced from the other. Now, the algebraic analogon of transforming $(\mathcal{T},\Theta_{\mathcal{T}})$ into a statistically equivalent staged tree is changing the order of summation (swapping) in eq. (4) or replacing products of indeterminates by lower degree monomials (resizing).

For instance, the nested factorisation for $(\mathcal{T},\Theta_{\mathcal{T}})$ from fig. 1 is given by

$$c_{\mathcal{T}}(\boldsymbol{\theta}) = \theta_0 + \theta_1(\theta_3+\theta_4) + \theta_2(\theta_3(\theta_5(\theta_7+\theta_8)+\theta_6)+\theta_4). \tag{5}$$

Note that the composition of edges in $\mathcal{T} = (V,E)$ can be read from this bracketing: inner sums of labels correspond to leaves in the tree, outer sums to root-labels. The tree $(\mathcal{S},\Theta_{\mathcal{S}})$ in fig. 2 is statistically equivalent to $(\mathcal{T},\Theta_{\mathcal{T}})$ via the following two polynomial transformations:

$$\begin{aligned} c_{\mathcal{T}}(\boldsymbol{\theta}) &= \theta_0 + \theta_3(\theta_1 + \theta_2(\theta_5(\theta_7+\theta_8)+\theta_6)) + \theta_4(\theta_1+\theta_2) && \text{(swap)} \\ &= \theta_0 + \theta_3(\theta_1 + \theta_2(\theta'_5+\theta''_5)+\theta_6)) + \theta_4(\theta_1+\theta_2) = c_{\mathcal{S}}(\boldsymbol{\theta}') && \text{(resize)} \end{aligned}$$

where $\boldsymbol{\theta}' = (\theta_0,\theta_1,\ldots,\theta_4,\theta'_5,\theta''_5,\theta_6)$ and $\theta'_5 = \theta_5\theta_7$, $\theta''_5 = \theta_5\theta_8$. Note that in $(\mathcal{T},\Theta_{\mathcal{T}})$, we have that $\theta_0+\theta_1+\theta_2 = 1$ whereas in $(\mathcal{S},\Theta_{\mathcal{S}})$ the values of these labels are renormalised to $\theta_0+\theta_1+\theta_3 = 1$ while retaining the global sum-to-1 condition on atomic probabilities.

Comparing figs. 1 and 2, we observe two important points: first, because both $(\mathcal{T},\Theta_{\mathcal{T}})$ and $(\mathcal{S},\Theta_{\mathcal{S}})$ represent the same model, and within this model the order of events 'move to one of the two cities' and 'attitude of landlord' is exchangeable. This is a very natural observation because we would not want to need to assume that the move of a student might cause a landlord's attitude or that these followed a particular chronological order. Second, the event $\Lambda(v_3)$ in $\mathcal{T}$ coincides with $\Lambda(w_3)$ in $\mathcal{S}$ but is depicted differently across both graphs. In particular, $\mathcal{T}$ illustrates this unfolding as two

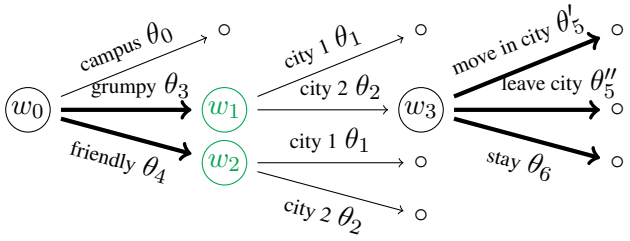

Figure 2: A staged tree $(\mathcal{S}, \Theta_{\mathcal{S}})$ which is represents the same model as $(\mathcal{T}, \Theta_{\mathcal{T}})$ from fig. 1. The thick edges correspond to the manipulated system when the event 'move to city 2' is enforced—they do not form a subtree anymore. See page 211 for an analysis.

individual decisions 'moving or not, deciding on location'. In contrast, $\mathcal{S}$ acknowledges that these are interlinked and depicts the one decision 'moving within city, moving out of the city, staying'. One can think of applications where one or the other depiction of this model is more expressive.

In a tree graph, for any fixed edge $e = (v, v')$ the event of passing through that edge $\Lambda(e) = \Lambda(v')$ is equal to the event of passing through its tail v'. Our causal manipulation above is an operation on vertices, whereas labels in staged trees are assigned to edges. We can thus translate algebraic or graphical projections or a differentiation operation of the interpolating polynomial to a vertex manipulation.

It has already been noted that *conditioning* operations on staged trees can be expressed in terms of projections of the underlying parameter space (Riccomagno and Smith, 2009). We will now show that a graphical projection from the idle $(\mathcal{T}, \Theta_{\mathcal{T}})$ onto some manipulated $(\mathcal{T}(v), \Theta_{\mathcal{T}(v)})$ where a unit was forced to pass through the fixed vertex v can be equivalently expressed as a local projection of the atomic probabilities combined with a differential operation.

Thus let

$$\mathfrak{d}_v(c_{\mathcal{T},A}(\boldsymbol{\theta})) = \frac{\partial^2 c_{\mathcal{T},A}(\boldsymbol{\theta})}{\partial\theta(e)\partial\mathbb{1}_{\Lambda(e)}}, \tag{6}$$

for the edge $e = (\cdot, v)$ with tail v and any $A \subseteq \Lambda(\mathcal{T})$, be a vertex-centred differentiation operation. By Görgen et al. (2015), $\mathfrak{d}_v(c_{\mathcal{T},A}(\boldsymbol{\theta}))$ simply equals the probability of A if A is an event depicted in the subtree $\mathcal{T}(v)$ and is zero otherwise. So $\mathfrak{d}_v$ performs Thwaites' vertex-intervention from eq. (2) in terms of a differential operation.

For instance, the interpolating polynomials in the example from page 209 for the idle $(\mathcal{T}, \Theta_{\mathcal{T}})$ and the manipulated staged tree $(\mathcal{T}(v_2), \Theta_{\mathcal{T}(v_2)})$ in fig. 1 equal

$$\begin{aligned} c_{\mathcal{T}}(\boldsymbol{\theta}) &= \theta_0 + \theta_1\theta_3 + \theta_1\theta_4 + \theta_2\theta_3\theta_5\theta_7 + \theta_2\theta_3\theta_5\theta_8 + \theta_2\theta_3\theta_6 + \theta_2\theta_4, \\ \mathfrak{d}_{v_2}(c_{\mathcal{T}}(\boldsymbol{\theta})) &= \frac{\partial}{\partial\theta_2} c_{\mathcal{T}}(\boldsymbol{\theta}) = \theta_3\theta_5\theta_7 + \theta_3\theta_5\theta_8 + \theta_3\theta_6 + \theta_4, \end{aligned} \tag{7}$$

respectively, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_8)$ and we omit the indicators for simplicity. If we are now interested in the probability of a student leaving the city of interest, assuming she was initially

forced to live there, we can calculate that

$$\pi_{\boldsymbol{\theta},\mathcal{T}}(\Lambda(v_4) \mid\mid \Lambda(v_2)) = \mathfrak{d}_{v_2}(c_{\mathcal{T},\Lambda(v_4)}(\boldsymbol{\theta})) = \frac{\partial^2 c_{\mathcal{T}(v_2)}(\boldsymbol{\theta})}{\partial\theta_7 \partial \mathbb{1}_{\Lambda(v_5)}} = \theta_3\theta_5 \tag{8}$$

where $e = (v_4, v_5)$ is the edge with label $\theta(e) = \theta_7$ that depicts the event 'leaving the city' in fig. 1. So this probability equals exactly the product of edge labels on the subpath from v_2 to v_4. It is thus simply the probability of passing from the root of the manipulated system to the vertex depicting the event of interest.

Let furthermore

$$\mathfrak{p}_v(c_{\mathcal{T},A}(\boldsymbol{\theta})) = c_{\mathcal{T}(v),A}(\boldsymbol{\theta}) \tag{9}$$

be the projection of a polynomial in $(\mathcal{T}, \Theta_{\mathcal{T}})$ to that polynomial in the subtree $(\mathcal{T}(v), \Theta_{\mathcal{T}(v)})$. Then we obtain:

Proposition 1 *Let $(\mathcal{T}, \Theta_{\mathcal{T}})$ be a staged tree, $\mathcal{T} = (V, E)$, and let $(\mathcal{T}(v), \Theta_{\mathcal{T}(v)})$ be the subtree after an intervention on $v \in V$ has taken place. Then the linear operator $\mathfrak{m}_v = \mathfrak{p}_v^{-1} \circ \mathfrak{d}_v$ performs a single vertex intervention and is commutative, so $\mathfrak{m}_v \circ \mathfrak{m}_w = \mathfrak{m}_w \circ \mathfrak{m}_v$ for $v, w \in V$.*

Proof We have seen in eq. (2) and the discussion of eq. (6) above that $\mathfrak{d}_v$ is a differential expression for forcing a unit to pass through the fixed vertex $v \in V$ as in Thwaites (2013). So $\mathfrak{d}_v$ performs a single-vertex intervention, eliminating all unfoldings in $(\mathcal{T}, \Theta_{\mathcal{T}})$ which are different from the ones depicted in $(\mathcal{T}(v), \Theta_{\mathcal{T}(v)})$, as well as the path leading from the root to v. Now $\mathfrak{p}_v^{-1} \circ \mathfrak{d}_v$ re-embeds this implicit projection into the larger system such that counterfactuals are not affected.

Following these steps, we immediately obtain

$$\mathfrak{m}_v = \mathfrak{p}_v^{-1} \circ \mathfrak{d}_v(c_{\mathcal{T},A}(\boldsymbol{\theta})) = \mathfrak{p}_v^{-1}(c_{\mathcal{T}(v),A}(\boldsymbol{\theta})) = c_{\mathcal{T}(v),A}(\boldsymbol{\theta}|_{\theta(\cdot,v)=1}) \tag{10}$$

so the composition $\mathfrak{m}_v$ effectively sets the edge-label leading into v to one, keeping the remainder of the staged tree invariant. This is a commutative operation. ■

The causal vertex-manipulation given by $\mathfrak{m}_v$ first differentiates with respect to the edge-label leading into v and then re-embeds the result into the idle system. This is important for applications where we would want to perform a sequence of interventions. By the proposition above, these can then be performed in any order. Note that we can interpret these calculations as a simple local differentiation operation

$$\mathfrak{m}_{v_j^\star}(c_{\mathcal{T}}(\boldsymbol{\theta})) = \sum_{(v_0,v_1)\in E(v_0)} \theta(v_0, v_1) \left(\cdots \left(\frac{\partial}{\partial \theta_j^\star} \sum_{(v_{j-1},v_j)\in E(v_j)} \theta(v_{j-1}, v_j) \left(\cdots \left(\sum_{(v_{k-1},v_k)\in E(v_{k-1})} \theta(v_{k-1}, v_k) \right) \right) \right) \right) \tag{11}$$

which affects only the relevant subtree. Of course, the ignorability of the respective local sum-to-one conditions is essential in this approach: clearly, a calculation like $\frac{\partial}{\partial\theta_1}\theta_1\theta_2$ would yield a very different result from a differentation of $\frac{\partial}{\partial\theta_1}\theta_1(1-\theta_1)$.

Note that if we were interested in manipulating all vertices in the same stage, thus forcing all units which arrive in a certain situation to pass on to a specified next situation, then this could be achieved by a composition of the relevant manipulations $\mathfrak{m}_{v_{i_k}} \circ \ldots \circ \mathfrak{m}_{v_{i_1}}$. This is then analogous to the BN type intervention on a product space.

So any intervention operation in staged trees relies only on a polynomial characterisation of the model in exactly the same way that the results on statistical equivalence classes of staged trees obtained by Görgen and Smith (2015) are encoded using the same polynomial representation. The great advantage of this algebraic or differential approach is its generality over a graph intervention.

For instance, the causal manipulation from page 209 of forcing a student to live in a certain city cannot be depicted as a subtree of $(\mathcal{S}, \Theta_{\mathcal{S}})$. It corresponds in fact to the thick edges in fig. 2 which are not connected. Thus, if we only knew about $(\mathcal{S}, \Theta_{\mathcal{S}})$ as a representation of our model—possibly inferred as the MAP staged tree from a dataset as in Cowell and Smith (2014) or from a consultation of domain experts—there would be no straight forward graphical way of expressing this manipulation on the idle system in terms of interventions on vertices. Similarly, or worse, if we were interested in assessing the effect of forcing students who are renting with a grumpy landlord to move, then $(\mathcal{T}, \Theta_{\mathcal{T}})$ could be easily manipulated to $(\mathcal{T}(v_4), \Theta_{\mathcal{T}(v_4)})$ but again $(\mathcal{S}, \Theta_{\mathcal{S}})$ would not allow us to answer this query because this time we would force a unit to go through two mutually exclusive edges simultaneously, following two different unfoldings from w_3.

So vertex-manipulations always act on a given graphical representation and might not be straightforward in different but statistically equivalent representations of the same class of staged trees. Our new differential approach however is independent of that graph representation and thus elegantly overcomes the restrictions noted above. Though of course our example setting is very simplified, these ideas extend to much more complex systems.

We note that our new operation allows also for efficient computations of the effect of an intervention. Because of the correspondence of a labelled tree graph $(\mathcal{T}, \Theta_{\mathcal{T}})$ with a nested expression of its interpolating polynomial $c_{\mathcal{T}}(\boldsymbol{\theta})$ as in eq. (4), the projection $\mathfrak{p}_v$ reduces a nested factorisation of the interpolating polynomial as in eq. (4) to an inner nesting, without violating the bracketing structure. For instance, we can easily check in eq. (7) that $c_{\mathcal{T}(v_2)}(\boldsymbol{\theta}) = \theta_3(\theta_5(\theta_7 + \theta_8) + \theta_6) + \theta_4$.

Lauritzen et al. (1990) found that this nested factorisation of a polynomial can lead to very fast algorithms that compute joint probabilities from marginals in BN models. This advantage is also enjoyed for staged trees, a much larger class of discrete parametric models, and can by the above thus be again used when calculating effects of a manipulation.

4. Discussion

We have been able to show in this work that within staged trees, causal manipulation is extremely simple: it is graphically very straight forward and thus easy communicate. And its effects can be easily calculated using a framework of differentiation and polynomial algebra which does not rely on a graphical representation. Our approach extends the usefulness of the ideas developed by Darwiche (2003) to causal inference, applicable both to BNs and much more general staged tree models. Hereby, the new operator we define can perform both Pearl's do-operation and Thwaites' vertex-centred manipulation while allowing for compositions of even more general types of interventions.

We are aware that because of this simplicity, our definition of a model in terms of a polynomial allows for even greater generality than already achieved. We can for instance develop the same differential theory as above, and define causal manipulation in these terms, without referring to a graphical representation of the model. These calculations are thus possible in any model whose atomic probabilities are defined in polynomial terms, and can be easily performed using computer algebra techniques even in large systems.

Acknowledgments

The first author is thankful to EPSRC for support (grant number EP/L505110/1).

References

L. M. Barclay, J. L. Hutton, and J. Q. Smith. Refining a Bayesian network using a Chain Event Graph. *Internat. J. Approx. Reason.*, 54(9):1300–1309, 2013.

H. Chan and A. Darwiche. When do numbers really matter? *J. Artificial Intelligence Res.*, 17: 265–287 (electronic), 2002.

R. A. Collazo and J. Q. Smith. A new family of Non-Local Priors for Chain Event Graph model selection. *Bayesian Analysis*, 2015.

R. G. Cowell and J. Q. Smith. Causal discovery through MAP selection of stratified Chain Event Graphs. *Electron. J. Stat.*, 8(1):965–997, 2014.

A. Darwiche. A differential approach to inference in Bayesian networks. *J. ACM*, 50(3):280–305 (electronic), 2003.

G. Freeman and J. Q. Smith. Bayesian MAP model selection of Chain Event Graphs. *J. Multivariate Anal.*, 102(7):1152–1165, 2011.

C. Görgen and J. Q. Smith. Equivalence Classes of Staged Trees. Preprint available from *arXiv:1512.00209v2 [math.ST]*, 2015.

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

S. L. Lauritzen, P. A. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990. Special issue on influence diagrams.

M. Leonelli, C. Görgen, and J. Q. Smith. Sensitivity analysis, multilinearity and beyond. Preprint available from *arXiv:1512.02266 [cs.AI]*, 2015.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995.

J. Pearl. *Causality*. Cambridge University Press, Cambridge, first edition, 2000. Models, reasoning, and inference.

E. Riccomagno and J. Q. Smith. The causal manipulation and Bayesian estimation of chain event graphs. *CRiSM*, (05-16), 2005.

E. Riccomagno and J. Q. Smith. The geometry of causal probability trees that are algebraically constrained. In *Optimal design and related areas in optimization and statistics*, volume 28 of *Springer Optim. Appl.*, pages 133–154. Springer, New York, 2009.

G. Shafer. *The Art of causal Conjecture*. MIT Press, Cambridge, 1996.

J. Q. Smith and P. E. Anderson. Conditional independence and Chain Event Graphs. *Artificial Intelligence*, 172(1):42–68, 2008.

P. Thwaites. Causal Identifiability via Chain Event Graphs. *Artificial Intelligence*, 195:291–315, 2013.

P. A. Thwaites, J. Q. Smith, and E. Riccomagno. Causal analysis with Chain Event Graphs. *Artificial Intelligence*, 174(12-13):889–909, 2010.