Tractable Bayesian Learning of Tree Belief Networks

Marina Meilă[1] Tommi Jaakkola[2]

Abstract

In this paper we present *decomposable priors*, a family of priors over structure and parameters of tree belief nets for which Bayesian learning with complete observations is tractable, in the sense that the posterior is also decomposable and can be completely determined analytically in polynomial time. Our result is the first where computing the normalization constant and averaging over a super-exponential number of graph structures can be performed in polynomial time. This follows from two main results: First, we show that factored distributions over spanning trees in a graph can be integrated in closed form. Second, we examine priors over tree parameters and show that a set of assumptions similar to (Heckerman and al., 1995) constrain the tree parameter priors to be a compactly parametrized product of Dirichlet distributions. Besides allowing for exact Bayesian learning, these results permit us to formulate a new class of tractable latent variable models in which the likelihood of a data point is computed through an ensemble average over tree structures.

1 Introduction

Effective inference in high-dimesional domains requires the combination of observed data with prior knowledge. When the prior knowledge is expressed as a probability distribution over the possible models and their parameters, Bayesian analysis provides the framework. In the field of graphical probability models, the first advances were made by [Spiegelhalter and Lauritzen, 1990, Dawid and Lauritzen, 1993] who considered priors and updates for a graphical model with a given, fixed, structure. In this context, the challenge is to define priors that can be expressed by a tractable number of hyper-parameter and for which computing the posterior is tractable.

Defining priors over both model structures and parameters, and effectively computing with them, is generally a much more challenging task. The reason is the very large (indeed super-exponential) number of possible model structures, each of them entailing a different parameter space. In order to circumvent this problem, several authors introduced independence assumptions that allow one to define the parameter prior by considering only the *local* graph structure (for example, a variable and its parents in a Bayes net, or a clique in a decomposable graphical model) [Dawid and Lauritzen, 1993, Heckerman et al., 1995, Giudici and Green, 1999, Madigan and Raftery, 1994]. The model classes most frequently analyzed are Bayesian networks [Cooper and Herskovits, 1992, Dawid and Lauritzen, 1993, Heckerman et al., 1995, Madigan and Raftery, 1994] and decomposable graphical models [Dawid and Lauritzen, 1993, Giudici and Green, 1999, Madigan and Raftery, 1994].

As prior for model structures [Giudici and Green, 1999, Madigan and Raftery, 1994, Cooper and Herskovits, 1992] use a uniform prior. Factored priors, suggested among others by [Madigan and Raftery, 1994] and used by [Heckerman et al., 1995], are priors where each edge of the graph con-

tributes a constant factor to the prior probability of each structure where it is present. Note that having a prior expressed in a product form over the graph edges does *not* imply independence between edges for either of these two classes. This point is discussed in detail in section 4. Note also that the above priors are only defined up to a normalization constant that cannot be computed in closed form, not even for the uniform prior of [Giudici and Green, 1999].

The above simplifications allow for a correspondingly factored form of the posterior if the observations are complete. In some special cases (e.g uninformative Dirichlet prior) the prior can be specified by a tractable number of hyper-parameters. In such cases the posterior probability of a model structure can be computed in closed form up to a multiplicative normalization constant and therefore the evidence for different model structures can be compared. Several authors developed search algorithms to find model structures with high posterior probability: [Cooper and Herskovits, 1992, Heckerman et al., 1995, Spirtes and Meek, 1995] for Bayesian networks, [Giudici and Green, 1999] for decomposable graphical models, and [Madigan and Raftery, 1994] for both. If model averaging is performed, this is done explicitly on a subset of the models examined during the search, typically the high scoring ones. It is not known how to find a set of models with provably high scores. Even finding the single model structure with highest posterior probability is intractable for Bayesian networks [Heckerman et al., 1995], polytrees [Dasgupta, 1999] and decomposable graphical models with bounded clique size [Srebro, 2001]. Averaging over all possible model structures in a familiy of non-polynomial size[3] is generally considered intractable and not attempted.

The family of graphical models whose structure is an undirected tree, called *trees* throughout this paper, was considered no exception.

In [Meilă and Jordan, 2000] and [Heckerman et al., 1995] (called HGC in the forthcoming) priors for trees were used with the assumption that computing the prior's normalization constant over the set of all trees is intractable. The present paper shows that this is not so: trees are the first super-exponential family of graphical models for which one can tractably integrate a distribution that factors over the edges.

The papers focuses on discrete variable domains, but our results also hold for real-valued variables with jointly normal distributions. We show that, with the standard assumptions of HGC (namely likelihood equivalence, parameter independence and parameter modularity), the prior for tree parameters is constrained to be a product of Dirichlet distributions whose hyper-parameters satisfy a set of consistency relations. We combine this with a factored prior for the tree structure to obtain what we call *decomposable* priors for structure and parameters. The term decomposable prior was first introduced by HGC in the context of general Bayes nets. It should not be confused with a *decomposable graphical model* [Pearl, 1988] which is a distribution over the variables. A decomposable prior over tree distributions can be represented with $\mathcal{O}(n^2)$ hyper-parameters and, more importantly, its normalization constant over tree structures and parameters can be computed analytically in closed form. This last result is a consequence of the fact that a factored distribution over tree structures can be integrated exactly, using a theorem from combinatorics called the *Matrix tree theorem*.

We show that if the prior is decomposable and we have a data set consisting of N complete i.i.d. observations, then the posterior distribution is also decomposable, and its hyper-parameters and normalization constant can be computed in $\mathcal{O}(n^3 + n^2 N)$ operations. Evaluating the posterior probability of a given tree takes then $\mathcal{O}(n)$ time.

The paper starts by defining tree distributions and the problem of Bayesian learning in section 2 and 3 respectively; it presents decomposable priors over tree structures and parameters in sections 4 and 5; the pieces of the puzzle are put together in section 6 where Bayesian learning for trees is described. Section 7 discusses possible extensions. In section 7.1 and 7.2 we present two cases of "disconnected" trees. In section 7.3 we exploit a different set of possibilities opened by our tractability results: we define a new model, *ensembles of trees*, and show that it can be learned by gradient ascent in the ML framework. Section 8 contains the final remarks.

2 Tree distributions

2.1 Tree distributions as graphical models

In this section we introduce the tree model and the notation that will be used throughout the paper. Let $V = \{1, \ldots, n\}$ denote the set of variables of interest. Let r_v be the number of values of variable $v \in V$, $r_{MAX} = \max r_v$, x_v a particular value of v, and x an assignment to all the variables in V.

According to the graphical model paradigm, each variable is viewed as a vertex of a graph. We shall call a graph that is connected and has no cycles a *tree* and shall denote by E its edge set. In this case E has exactly $n-1$ edges. In this definition, we differ from the traditional graphical models terminology (see [Pearl, 1988] or [Meilă and Jordan, 2000]) where trees and polytrees are not required to be connected. We will discuss the case of disconnected trees in section 7.

Now we define a probability distribution T that is *factored according*

to a tree. Let us denote by T_{uv} and T_v the marginals of T:

$$\begin{aligned} T_{uv}(x_u, x_v) &= \sum_{x:u=x_u, v=x_v} T(x) \\ T_v(x_v) &= \sum_{x:v=x_v} T(x). \end{aligned}$$

Let $\deg v$ be the *degree* of vertex v, i.e. the number of edges incident to $v \in V$. Then, the distribution T is factored according to the tree (V, E) if it can be represented as:

$$T(x) = \prod_{v \in V} T_v(x_v) \prod_{uv \in E} \frac{T_{uv}(x_u, x_v)}{T_u(x_u) T_v(x_v)} \tag{1}$$

The distribution itself will be called a tree when no confusion is possible.

In the above definition, $T(x)$ factors into a product of variable marginals, which depends only on the variable set V, and a product over the edges $uv \in E$, in which each factor depends only on the respective edge marginal T_{uv}. This factors is symmetric in u, v, reflecting the fact that the edges E of the graph are undirected.

A tree (V, E) can be transformed into a *directed tree* by choosing an arbitrary node as *root*. The edges are recursively directed outwards from the root. We shall denote a thus directed edge set with $\overline{E}$. For the example in figure 1 the undirected edge set is $E = \{ab, bc, bd\}$; if node a is chosen as root, the directed edge set is $\overline{E} = \{\overline{ab}, \overline{bc}, \overline{bd}\}$ (figure 1,b) and if b is chosen as root then we obtain $\overline{E} = \{\overline{ba}, \overline{bc}, \overline{bd}\}$. If a directed edge $\overline{uv}$ goes from u to v we say that u is the *parent* of v (and respectively that v is the *child* of u). The above procedure creates a directed graph where each node has at most one parent (the root being the only node with no parents).

A distribution that factors according to the tree (V, E) by equation (1) can be put in a new factored form that matches a given directed tree

$(V, \overline{E})$ obtained from (V, E). To do this, one applies the substitution

$$\frac{T_{vu}}{T_u} \;=\; T_{v|u}$$

for each directed edge $\overline{uv}$ in $\overline{E}$. In [Meilă and Jordan, 2000] it is shown that after simplifying the common factors, the denominator of equation (1) contains exactly one T_v factor for each child of v. The result is called the *directed representation* of the tree distribution:

$$T(x) \;=\; \prod_{v \in V} T_{v|\mathrm{pa}(v)}(x_v|x_{\mathrm{pa}(v)}) \tag{2}$$

where $\mathrm{pa}(v)$ represents the parent of v in the thus directed tree or the empty set if v is the root. By operating the reverse substition one obtains the undirected representation from the directed one.

There is one distinct form (2) for each choice of the root node. The undirected tree representation (1) and the directed representations (2) are equivalent, in the sense that for any variable configuration x we obtain the same value of the probability $T(x)$ no matter which representation is used to compute it. For an extensive presentation of tree distributions and their properties the reader is referred to [Meilă and Jordan, 2000]. Here we only add that both representations have advantages that will be drawn upon in the next sections: the directed tree representation exhibits the parameter independencies, while the undirected representation, due to its symmetry, will be most useful otherwise.

A tree distribution is a graphical model. In this context, the underlying (undirected) graph is called the *structure* of the tree graphical model. Since in this paper the focus is on the family of trees which share the same set of variables V but have different sets of edges, we will assimilate the structure of a tree with its edge set E and henceforth we will call E the tree structure.

Tree graphical models are a subset of the more general class of *polytrees* [Pearl, 1988]. While in a tree a node always has at most one parent, in polytrees nodes can have more than one parent. For the results in this paper it is essential that each node has no more than one parent. Therefore they do not generalize to polytrees. Trees are also a subclass of Bayes nets, Markov nets and decomposable models (see e.g. [Pearl, 1988]). The results in this paper do not extend to these model classes.

2.2 Parametrizations

A tree distribution is described by its structure and parameters. In the following, without loss of generality, we will consider that both the directed and the undirected tree representation are in the *probability table* parametrizations. For a tree distribution T with structure E over the set of variables V we define

$$\begin{aligned} \theta_E &= \{\theta_{uv}(ij),\, uv \in E,\, i = 1, \ldots r_u,\, j = 1, \ldots r_v\} \\ \theta_{uv}(ij) &= T_{uv}(ij) \qquad (3) \\ \theta_v(j) &= T_v(j) \qquad (4) \end{aligned}$$

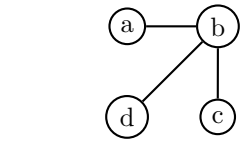
In the directed representation corresponding to an orientation $\overline{E}$ of E the same distribution is described by the parameters

$$\begin{aligned} \theta_{\overline{E}} &= \{\theta_{u|v}(ij),\, \overline{vu} \in \overline{E},\, i = 1, \ldots r_u,\, j = 1, \ldots r_v\} \\ \theta_{u|v}(i|j) &= T_{u|v}(i|j) \qquad (5) \end{aligned}$$

To simplify notation we assume by convention that if $v \in V$ is the root, then $\mathrm{pa}(v)$ takes one value only and $\theta_{v|\mathrm{pa}(v)} = \theta_v$.

Hence, each directed representation of T has a distinct set of parameters $\theta_{\overline{E}}$. These parameters are related to the parameters θ_E of

i.



$$T = \frac{T_{ab}}{T_a T_b}\frac{T_{bc}}{T_b T_c}\frac{T_{bd}}{T_b T_d} T_a T_b T_c T_d$$

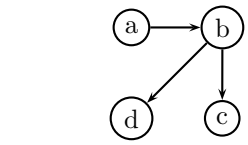
$\{\theta_{ab}(00)\ \theta_{ab}(01)\ \theta_{ab}(10)\ \theta_{ab}(11)\}$ $\{\theta_a(0)\ \theta_a(1)\}$

$\{\theta_{bc}(00)\ \theta_{bc}(01)\ \theta_{bc}(10)\ \theta_{bc}(11))\}$ $\{\theta_b(0)\ \theta_b(1)\}$

$\{\theta_{bd}(00)\ \theta_{bd}(01)\ \theta_{bd}(10)\ \theta_{bd}(11)\}$ $\{\theta_c(0)\ \theta_c(1)\}$

$\{\theta_d(0)\ \theta_d(1)\}$

ii.



$$T = T_a T_{b|a} T_{c|b} T_{d|b}$$

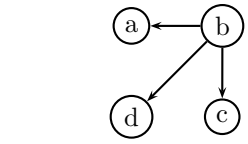
$\{\theta_a(0)\ \theta_a(1)\}$

$\{\theta_{b|a}(0|0)\ \theta_{b|a}(1|0)\}$ $\{\theta_{b|a}(0|1)\ \theta_{b|a}(1|1)\}$

$\{\theta_{c|b}(0|0)\ \theta_{c|b}(1|0)\}$ $\{\theta_{c|b}(0|1)\ \theta_{c|b}(1|1)\}$

$\{\theta_{d|b}(0|0)\ \theta_{d|b}(1|0)\}$ $\{\theta_{d|b}(0|1)\ \theta_{d|b}(1|1)\}$

iii.



$$T = T_b T_{a|b} T_{c|b} T_{d|b}$$

$\{\theta_b(0)\ \theta_b(1)\}$

$\{\theta_{a|b}(0|0)\ \theta_{a|b}(1|0)\}$ $\{\theta_{a|b}(0|1)\ \theta_{a|b}(1|1)\}$

$\{\theta_{c|b}(0|0)\ \theta_{c|b}(1|0)\}$ $\{\theta_{c|b}(0|1)\ \theta_{c|b}(1|1)\}$

$\{\theta_{d|b}(0|0)\ \theta_{d|b}(1|0)\}$ $\{\theta_{d|b}(0|1)\ \theta_{d|b}(1|1)\}$

Figure 1: A tree over four variables in three equivalent representations: **i.** undirected, **ii.** directed with root a and **iii.** directed with root b. The parameters for each representation are enumerated next to the graph, assuming that a, b, c, d take values in $\{0, 1\}$. The groups of parameters that sum to 1 are enclosed in braces. The sets of parameters in (i) and (ii) are related by the equations: $\theta_{b|a}(00) = \frac{\theta_{ab}(00)}{\theta_a(0)}$ ..... $\theta_{d|b}(11) = \frac{\theta_{bd}(11)}{\theta_b(1)}$ The total number of free parameters for this tree is equal to 7, verifying (10).

the undirected representation by

$$\theta_{u|v}(i|j) \;=\; \frac{\theta_{uv}(ij)}{\theta_v(j)} \qquad \text{for } \overline{vu} \in \overline{E},\; i = 1, \ldots r_u,\; j = 1, \ldots r_v \tag{6}$$

In addition, the parameters satisfy the usual normalization constraints

$$\sum_{i=1}^{r_u} \theta_{u|v}(i|j) \;=\; 1 \qquad \text{for } \overline{vu} \in \overline{E},\; j = 1, \ldots r_v \tag{7}$$

$$\sum_{i=1}^{r_u} \theta_{uv}(ij) \;=\; \theta_v(j) \qquad \text{for } uv \in E,\; j = 1, \ldots r_v \tag{8}$$

$$\sum_{j=1}^{r_v} \theta_v(j) \;=\; 1 \qquad \text{for } v \in V \tag{9}$$

Taking into account the normalization constraints (7–9), the total number of free parameters in either representation equals

$$\sum_{uv \in E} r_u r_v - \sum_{v \in V} (\deg v - 1) r_v - 1 \tag{10}$$

In the framework of graphical models, trees stand out by their special computational advantages (presented in detail in [Meilă-Predoviciu, 1999]). Inference and sampling from a tree are linear in the number of variables n. Finding the Maximum Likelihood (ML) tree structure and parameters over n discrete variables can be done in quadratic time by an algorithm due to [Chow and Liu, 1968]. This was generalized to Maximum A-Posteriori (MAP) estimation in [Meilă-Predoviciu, 1999, Heckerman et al., 1995]. In the following sections we present another remarkable property of tree graphical models: the fact that, under standard assumptions, computing the exact posterior and averaging under it (something we will call Bayesian learning in the forthcoming) is tractable.

3 The Bayesian learning problem

We now turn to the problem of learning trees in the Bayesian framework. In this framework, one assumes a prior $P_0(T)$ over the set $\mathcal{T}_V$ of all tree distributions defined on the domain V. Learning from a dataset of complete and independently generated observations $\mathcal{D} = \{x^1, x^2, \ldots x^N\}$ means finding the *posterior* distribution $P(T|\mathcal{D})$ over the set of models $\mathcal{T}_V$. The solution to this problem is given by the well known Bayes' formula

$$P(T|\mathcal{D}) \;=\; \frac{P_0(T)\prod_{t=1}^{N} T(x^t)}{P(\mathcal{D})} \tag{11}$$

Practically however, Bayesian leaning poses a number of significant challenges. First, one needs to define a distribution over the space of all models to play the role of the prior. Such a distribution is composed of a discrete distribution over the set of tree structures $P_0(E)$ and a probability density over the continuous set of tree parameters. For instance, for the undirected parametrization θ_E, the prior can be written as

$$P_0(T) \;=\; P_0(E)P_0(\theta_E|E) \tag{12}$$

Since for each tree structure E there are several equivalent parametrizations, we must define a prior that changes consistently from one set of parameters to another. Moreover, the second factor in the above formula requires us to define a prior distribution for the tree parameters for each possible structure E. The discrete space of all spanning tree structures over V has a super-exponential number of trees (n^{n-2}) [West, 1996] which makes defining a distribution over it a non-trivial task. Thus, the first practical requirement is to have a tractable representation for the prior.

Second, even with a tractable representation, the explicit computation of the posterior $P(T|\mathcal{D})$ is usually intractable due to the dif-

ficulty of computing the normalization constant $P(\mathcal{D})$ in (11). This is shown, for instance, in HGC, where in some cases the prior can be tractably represented but the normalization constant of the posterior is intractable. Common practices in Bayesian learning are Maximum A-Posteriori (MAP) estimation [Meilă and Jordan, 2000], approximations of the posterior around its peaks [Cheeseman and Stutz, 1995, Heckerman et al., 1995, Madigan and Raftery, 1994] or Markov Chain Monte Carlo [Dellaportas and Forster, 1999, Giudici and Green, 1999]. An exception from this is taken by the so-called *conjugate priors*. If a given (graphical) model has a family of conjugate priors $\mathcal{P}$ then for $P_0 \in \mathcal{P}$ the posterior is also in $\mathcal{P}$. The property of having conjugate priors is characteristic of the exponential family of distributions [DeGroot, 1975]. In this paper we set out to find the conjugate prior for the family of spanning tree models $\mathcal{T}_V$.

According to (12), to define a prior over $\mathcal{T}_V$ one needs to define a prior for tree structures and a prior for parameters, given the structure. While it is not hard to see that for a fixed structure E a tree distribution over discrete variables is an exponential model and thus has conjugate priors [Dawid and Lauritzen, 1993], realizing the same fact when E also varies is by far less obvious and constitutes the main contribution of this paper. In the next section we establish the core theorem that allows us to do so.

4 Decomposable distributions over tree structures

A *decomposable* distribution P over spanning tree structures E can be defined by a set of hyper-parameters $\beta_{uv} = \beta_{vu} \geq 0$; $\beta_{vv} = 0$, $u, v \in V$

by

$$P(E) \;=\; \frac{1}{Z} \prod_{uv \in E} \beta_{uv}. \tag{13}$$

In the above, Z is the normalization constant

$$Z \;=\; \sum_{E} \prod_{uv \in E} \beta_{uv}. \tag{14}$$

Note that in the distribution (14), each hyper-parameter β_{uv} can be interpreted as the weight of edge uv, and the probability of a structure E is the product of the weights of all edges in E.

Although this distribution is expressible in a product form, it *does not* imply that the edges' occurrences in E are independent, since the set E as a whole is constrained to be a tree structure. To see this, take for example the domain $V = \{a, b, c, d\}$ with $\beta_{uv} = 1$ for all $u, v \in V$, $u \neq v$. Clearly, if $ac, bc \in E$ then $ab \notin E$ therefore $P(ab \,|\, ac, bc) = 0$. However, $P(ab) > 0$ as there obviously are possible tree structures that contain edge ab. Thus, the probability of any edge ab is dependent on knowledge about the presence of other edges; this is due to the constraint that the set of edges represented by E contains no cycles. Compare also with the Dirichlet distribution (defined in the next section): under the Dirichlet distribution the variables are not independent although the distribution has a product form

This prior is simple and compactly parametrized, but to be completely defined one needs to evaluate the normalization constant Z. Computing Z the direct way by using formula (14) is intractable, since one needs to sum over n^{n-2} terms. However, the following theorem shows a practical and exact method for doing so.

Let us start by introducing a simplifying notation to refer to a set of real values each corresponding to a pair of variables in V

$$a \;=\; \{a_{uv},\, u, v \in V, u \neq v\} \tag{15}$$

In addition, $a \geq 0$ will mean that $a_{uv} \geq 0$, $a_{uv} \in a$; the product ab will denote $\{a_{uv}b_{uv},\, u, v \in V, u \neq v\}$ for a, b defined as above. With the new notation, we can say that a decomposable distribution $P(E)$ is defined by a set of hyper-parameters β.

Theorem 1 *Let $P(E)$ be a distribution over spanning tree structures defined by (13,14). Then the normalization constant Z is equal to the determinant $|Q(\beta)|$, with $Q(\beta)$ representing the first $(n-1)$ rows and columns of the matrix $\overline{Q}(\beta)$ given by:*

$$\overline{Q}_{uv}(\beta) = \overline{Q}_{vu}(\beta) = \begin{cases} -\beta_{uv} & 1 \leq u < v \leq n \\ \sum_{v'=1}^{n} \beta_{v'v} & 1 \leq u = v \leq n \end{cases} \tag{16}$$

This shows that summing over the distribution of all trees, when this distribution factors according to the trees' edges, can be done in closed form by computing the value of an order $n-1$ determinant. This takes $\mathcal{O}(n^3)$ operations. The theorem is a generalization to real-valued weights of a remarkable result in graph theory called the matrix tree theorem [West, 1996]. The matrix tree theorem, the proof of theorem 1 as well as the other proofs appear in the appendix.

In the following it will be useful to think of $\overline{Q}(\beta)$ and $Q(\beta)$ as functions mapping a set of parameters β each corresponding to a pair of variables in V to a matrix in the ways described by theorem 1. Note also that $|\overline{Q}(\beta)| = 0$ as the rows (columns) of matrix $\overline{Q}(\beta)$ sum to 0.

Examples. The *uniform* distribution given by $\beta_{uv} = 1$ is decomposable. Its normalization constant is $Z = n^{n-2}$. Hence $P(E) = \frac{1}{n^{n-2}}$ for all spanning tree structures E.

Another example is the distribution given by

$$\beta_{uv} \;=\; \begin{cases} 1 & uv \in E^* \\ \beta < 1 & \text{otherwise} \end{cases} \tag{17}$$

The probability of a tree structure under this distribution is given by $P(E) \propto \beta^{|E \setminus E^*|}$. We can interpret this as penalizing a tree by a factor

β for each edge that is not contained in some desired set of edges E^*. In particular, if E^* itself is a spanning tree, then the probability of a tree structure E decays exponentially with the number of differences from the "gold standard" E^*. This prior was suggested in HGC in the context of learning of directed graphical models.

The support graph. The factored form of the decomposable distribution makes it easy to test whether a given structure has non-zero probability. If all the β hyper-parameters are strictly positive, then every tree structure is possible. Otherwise, the structures that will never appear are the structures containing one or more zero-weight edges. We denote by E^{sup} the set of edges uv for which $\beta_{uv} > 0$. The graph $G^{sup} = (V, E^{sup})$ is called *the support graph* of $P(E)$. If enough edges have zero weights, then G^{sup} may be disconnected. In the following we shall assume that the support graph is connected, leaving the discussion of the general case for section 7.1.

In the remainder of this section we develop a number of consequences of theorem 1.

Computing averages under a decomposable distribution A decomposable distribution is a (curved) exponential model [Murray and Rice, 1993] and $\ln Z$ represents its *cumulant generating function* or *partition function*. Many quantities of interest, like averages under $P(E)$ can be expressed as derivatives of the partition function. The next series of results exemplifies these possibilities. We assume that G^{sup} is connected.

Lemma 2 *Let Z be given by equation (14) with $\beta \geq 0$ and G^{sup} connected. Then the partial derivative of Z with respect to β_{uv} is*

$$\frac{\partial Z}{\partial \beta_{uv}} \;=\; M_{uv}(\beta)|Q(\beta)|. \tag{18}$$

where $Q(\beta)$ is given by theorem 1, Q^{-1} is the inverse of Q and $M(\beta)$

is a symmetric matrix with 0 diagonal defined by

$$\begin{aligned} M_{uv} &= (Q^{-1})_{uu} + (Q^{-1})_{vv} - 2(Q^{-1})_{uv},\ u, v < n \\ M_{nv} &= M_{vn} = (Q^{-1})_{vv},\ v < n \\ M_{vv} &= 0 \end{aligned} \tag{19}$$

We shall denote by $< f >_P$ the average of a function f under a distribution P. The following lemma states a useful fact about averages of additive functions. An *additive* function $f(E)$ satisfies

$$f(E) = \sum_{uv \in E} f_{uv} \tag{20}$$

for any spanning tree structure E.

Lemma 3 *Let $P(E)$, Q and M be given by (13), theorem 1 and (19) respectively and f be an additive function of the structure E. Then the average of f under P is*

$$< f(E) >_P = \sum_E f(E)P(E) \tag{21}$$

$$= \sum_{u<v} f_{uv}\beta_{uv}M_{uv}(\beta) \tag{22}$$

$$= \text{trace}\,[\,Q(\beta f)Q^{-1}(\beta)\,] \tag{23}$$

In (23), f is an overloaded notation representing the set $\{f_{uv},\ u, v \in V\}$ in the sense of (15). A similar but more obvious result holds for functions $g(E)$ that are *multiplicative*, i.e. $g(E) = \prod_{uv \in E} g_{uv}$. For such functions we obtain

$$< g(E) >_P = \frac{|Q(\beta g)|}{|Q(\beta)|} = |Q(\beta g)Q^{-1}(\beta)| \tag{24}$$

Note that the likelihood $T(x)$ is a multiplicative function and its logarithm is therefore additive. Hence the above lemmas can be applied to compute partial derivatives of the likelihood w.r.t model parameters for instance. We will make use of equation 24 in section 6.

5 Decomposable priors over tree parameters

5.1 Assumptions

Now we examine priors over tree parameters, with the goal of finding conditions under which the priors can be tractably represented. These conditions will take the form of a series of restrictive assumptions about the prior.

First let us keep the distribution T fixed. As shown in section 2 this distribution can be represented either by (1) or by (2), the latter representation having a distinct form for each possible choice of the root(s). These representations however will assign exactly the same probability $T(x)$ to an observation x, so there is no way to distinguish between them from the point of view of the data. Thus we shall require that the corresponding parameter sets are also the same from the point of view of the prior. This leads to the assumption of *Likelihood equivalence*:

Assumption 1 (Likelihood equivalence) *Let T be a tree distribution having structure E, $\overline{E}$ a directed tree structure obtained from E and $\theta_E, \theta_{\overline{E}}$ the respective parameters of T. Denote by $abs\left|\frac{\partial\theta_{\overline{E}}}{\partial\theta_E}\right|$ the magnitude of the Jacobian of the transformation $\theta_E \rightarrow \theta_{\overline{E}}$. Then*

$$P_0(\theta_{\overline{E}}(\theta_E)|\overline{E})\, abs\left|\frac{\partial\theta_{\overline{E}}}{\partial\theta_E}\right| = P_0(\theta_E|E) \tag{25}$$

This assumption states that in all possible parametrizations consistent with a given structure E the prior will assign the same probability mass to any given (measurable) subset in parameter space. Thus, the prior treats likelihood equivalent parametrizations as indistinguishable. Assumption 1 also allows us to use in the future whichever of $P_0(\theta_E|E)$

and $P_0(\theta_{\overline{E}}|\overline{E})$ is most convenient, since we can always obtain one from the other via equation (25).

A simple example will help illustrate this assumption. Let us consider the "tree" over two variables $V = \{u, v\}$, $E = \{uv\}$. The undirected parametrization of this tree is given by

$$\theta_E \;=\; \{\theta_{uv}(00),\; \theta_{uv}(01),\; \theta_{uv}(10)\}$$

The directed parametrization corresponding to $\overline{E} = \{\overline{uv}\}$ is

$$\theta_{\overline{E}} \;=\; \{\theta_u(0),\; \theta_{v|u}(0|0),\; \theta_{v|u}(0|1)\}$$

In the above equation we have included in θ_E and $\theta_{\overline{E}}$ only the free parameters, in order to ensure that the Jacobian matrix is square and non-singular. The mapping between the two sets of parameters is:

$$\theta_u(0) \;=\; \theta_{uv}(00) + \theta_{uv}(01) \tag{26}$$

$$\theta_{v|u}(0|0) \;=\; \frac{\theta_{uv}(00)}{\theta_{uv}(00) + \theta_{uv}(01)} \tag{27}$$

$$\theta_{v|u}(0|1) \;=\; \frac{\theta_{uv}(10)}{1 - \theta_{uv}(00) - \theta_{uv}(01)} \tag{28}$$

From the mapping we can compute the Jacobian

$$\left|\frac{\partial\theta_{\overline{E}}}{\partial\theta_E}\right| \;=\; \left|\begin{array}{ccc} \frac{\partial\theta_u(0)}{\partial\theta_{uv}(00)} & \frac{\partial\theta_u(0)}{\partial\theta_{uv}(01)} & \frac{\partial\theta_u(0)}{\partial\theta_{uv}(10)} \\ \frac{\partial\theta_{v|u}(0|0)}{\partial\theta_{uv}(00)} & \frac{\partial\theta_{v|u}(0|0)}{\partial\theta_{uv}(01)} & \frac{\partial\theta_{v|u}(0|0)}{\partial\theta_{uv}(10)} \\ \frac{\partial\theta_{v|u}(0|1)}{\partial\theta_{uv}(00)} & \frac{\partial\theta_{v|u}(0|1)}{\partial\theta_{uv}(01)} & \frac{\partial\theta_{v|u}(0|1)}{\partial\theta_{uv}(10)} \end{array}\right|$$

After performing the calculations, the Jacobian becomes

$$\left|\frac{\partial\theta_{\overline{E}}}{\partial\theta_E}\right| \;=\; \frac{1}{(\theta_{uv}(00) + \theta_{uv}(01))(1 - \theta_{uv}(00) - \theta_{uv}(01))} \;=\; -\frac{1}{\theta_u(0)\theta_u(1)} \tag{29}$$

Note that in this case the Jacobian is negative, therefore using its absolute value in equation (25) is necessary.

If we denote by $\overline{E}' = \{\overline{vu}\}$ the opposite orientation, then by symmetry we have

$$\left|\frac{\partial\theta_{\overline{E}}}{\partial\theta_E}\right| \;=\; -\frac{1}{\theta_v(0)\theta_v(1)}$$

Note also that

$$\int P_0(\theta_E)d\theta_E \;=\; \int P_0(\theta_{\overline{E}})d\theta_{\overline{E}} \;=\; \int P_0(\theta_{\overline{E}'})d\theta_{\overline{E}'} \;=\; 1$$

Thus, by assuming likelihood equivalence, we effectively recognize that each tree distribution is an equivalence class containing all its different parametrizations.

Likelihood equivalence has the effect of compressing the space that we have to define P_0 on, but it still leaves us with the task of assigning a separate prior for the parameters of each (undirected) tree structure. We now transform this problem into one of assigning a prior for each of the possible tree edges by making the following additional assumptions:

Assumption 2 (Parameter independence) *For any structure* $\overline{E}$ *and any* $\overline{vu} \in \overline{E},\ j, j' = 1, \ldots r_v,\ j' \neq j$ *the parameter vectors* $\theta_{u|v}(.|j)$ *and* $\theta_{u|v}(.|j')$ *are independent under* P_0*. The parameters* $\theta_{u|v}(.|j)$ *are also independent under* P_0 *of the parameter sets* $\theta_{u'|v'}(.|j')$ *corresponding to any other edge in* $\overline{E}$*.*

Assumption 3 (Parameter modularity) *The prior* $P_0(\theta_{u|v}|\overline{E})$ *is the same for all structures* $\overline{E}$ *that contain the edge* $\overline{vu}$*.*

In other words, parameter independence states that the prior over parameters factors into a product over the edges.

$$P_0(\theta_{\overline{E}}\,|\,\overline{E}) \;=\; \prod_{v \in V} \prod_{j=1}^{r_{\mathrm{pa}(v)}} P_0(\theta_{v|\mathrm{pa}(v)}(.|j)\,|\,\overline{E}) \tag{30}$$

By stating in addition that the prior for an edge is the same for all tree structures that contain that edge, we have effectively removed the dependence on $\overline{E}$ (or E) from the parameters prior. Therefore, instead of having to define a separate prior $P_0(\theta_{\overline{E}}\,|\,\overline{E})$ for each possible tree structure $\overline{E}$, with the previous three assumptions we need only define the pairwise priors $P_0(\theta_{uv}(.,.))$ for u, v in V in order to have priors

for all possible sets of parameters $\theta_{\overline{E}}$. From now on, we will write $P_0(\theta_E)$, $P_0(\theta_{\overline{E}})$ instead of $P_0(\theta_E|E)$ and $P_0(\theta_{\overline{E}}|\overline{E})$ respectively. It is convenient to define the prior in terms of the undirected parameters θ_{uv}, $u, v \in V$. From it we can obtain the prior $P_0(\theta_{\overline{E}}|\overline{E})$ for any directed set of parameters $\theta_{\overline{E}}$ via equation (25).

We shall call a prior P_0 satisfying assumptions 1, 2 and 3 a *decomposable prior* for tree parameters. If both $P_0(E)$ and $P_0(\theta_E)$ are decomposable, the resulting prior over tree distribution is also called *decomposable*. For now we also assume that G^{sup} is connected.

Assumption 4 (Connectivity) *The support graph of $P_0(E)$ is connected.*

5.2 The Dirichlet prior

As we shall see now, the assumptions we made also constrain the functional form the prior can have.

Theorem 4 *Let $P(T) = P(E)P(\theta_E)$ be a decomposable distribution over tree parameters, for which the support graph of $P(E)$ is connected and $P(\theta_E) > 0$ for $\theta_E > 0$. Then for any tree T in any directed representation $\overline{E}, \theta_{\overline{E}}$:*

$$P(\theta_{\overline{E}}|\overline{E}) = \prod_{v \in V} P(\theta_{v|\mathrm{pa}(v)}) \tag{31}$$

$$P(\theta_{v|u}) = \prod_{i=1}^{r_u} D(\theta_{v|u}(.|i)\,;\, N'_{vu}(.i)) \tag{32}$$

where D is the Dirichlet distribution and $N'_{vu}(ij) > 0$ are its hyperparameters. The numbers $N'_{uv}(ij) = N'_{vu}(ji)$ are defined for all edges uv with $\beta_{uv} > 0$ and satisfy

$$\sum_{i=1}^{r_u} N'_{uv}(ij) = N'_v(j) \tag{33}$$

$$\sum_{j=1}^{r_v} N'_v(j) \;=\; N' \tag{34}$$

The Dirichlet distribution [DeGroot, 1975] is defined over the $\{\theta_1, \ldots \theta_r, \mid \sum_j \theta_j = 1,\, \theta_j > 0, j = 1, \ldots r\}$ by

$$D(\theta_1, \ldots \theta_r; N'_1, \ldots N'_r) \;=\; \frac{1}{Z_D} \prod_{j=1}^{r} \theta_j^{N'_j - 1} \tag{35}$$

The numbers $N'_1, \ldots N'_r > 0$ are the hyper-parameters of the Dirichlet prior; their sum is denoted by N'. The normalization constant Z_D has the form

$$Z_D \;=\; \frac{\prod_{j=1}^{r} \Gamma(N'_j)}{\Gamma(N')}$$

with Γ denoting the Euler function $\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx$.

For the previous two variables example and $\overline{E} = \{\overline{uv}\}$ as before, we have

$$\begin{aligned}
P_0(\theta_{\overline{E}}) \;=& \\
&= D(\theta_u(0), \theta_u(1)\,;\, N'_u(0), N'_u(1)) \cdot D(\theta_{v|u}(0|0), \theta_{v|u}(1|0)\,;\, N'_{uv}(00), N'_{uv}(10)) \\
&\quad \cdot D(\theta_{v|u}(0|1), \theta_{v|u}(1|1)\,;\, N'_{uv}(01), N'_{uv}(11)) \\
&= \frac{\Gamma(N')}{\Gamma[N'_u(0)]\Gamma[N'_u(1)]} \theta_u(0)^{N'_u(0)-1} \theta_u(1)^{N'_u(1)-1} \\
&\quad \cdot \frac{\Gamma[N'_u(0)]}{\Gamma[N'_{uv}(00)]\Gamma[N'_{uv}(01)]} \theta_{v|u}(0|0)^{N'_{uv}(00)-1} \theta_{v|u}(1|0)^{N'_{uv}(01)-1} \\
&\quad \cdot \frac{\Gamma[N'_u(1)]}{\Gamma[N'_{uv}(10)]\Gamma[N'_{uv}(11)]} \theta_{v|u}(0|1)^{N'_{uv}(10)-1} \theta_{v|u}(1|1)^{N'_{uv}(11)-1}
\end{aligned}$$

Replacing now the parameters $\theta_{\overline{E}}$ by their expressions in equations (26–28) and performing the calculations we obtain

$$\begin{aligned}
P_0(\theta_{\overline{E}}) \;=&\; \underbrace{\theta_u(0)\theta_u(1)}_{\left|\frac{\partial \theta_{\overline{E}}}{\partial \theta_E}\right|^{-1}} \\
=&\; \underbrace{\cdot \frac{\Gamma(N')\theta_{uv}(00)^{N'_{uv}(00)-1}\theta_{uv}(01)^{N'_{uv}(01)-1}\theta_{uv}(10)^{N'_{uv}(10)-1}\theta_{uv}(11)^{N'_{uv}(11)-1}}{\Gamma[N'_{uv}(10)]\Gamma[N'_{uv}(11)]\Gamma[N'_{uv}(00)]\Gamma[N'_{uv}(01)]}}_{D(\theta_{uv}(00),\theta_{uv}(01),\theta_{uv}(10),\theta_{uv}(11)\,;\, N'_{uv}(00),N'_{uv}(01),N'_{uv}(10),N'_{uv}(11)}
\end{aligned}$$

This equation agrees with the likelihood equivalence assumption and confirms that if the prior for a directed representation is Dirichlet, then the prior for the undirected representation is also Dirichlet.

5.3 Discussion

The above line of reasoning parallels the one in HGC, where the Dirichlet prior for general Bayes nets was derived from assumptions similar to 1–3. But, unlike the case of general Bayes nets, where the prior is specified by an exponential number of hyper-parameters, in the case of tree graphical models the prior can be described by a set of only $\mathcal{O}(n^2 r_{MAX}^2)$ "pairwise marginal counts" $N'_{uv}(ij)$. This is possible because in the space of tree structures the likelihood equivalence classes can be explicitly represented [4] and the number of possible parents for a variable is no larger than one. Therefore, not only the tree belief net itself, but also any decomposable distribution over trees can be completely defined in terms of pairwise interactions.

Another technical difference is that, in HGC, one key part of the framework is the *complete model* represented by the graph with no missing edges. The complete graph being obviously not in $\mathcal{T}_V$, this paper reconstructs the framework without recourse to it.

For each fixed tree structure, our prior is strongly hyper-Markov [Dawid and Lauritzen, 1993] (in fact it is a hyper-Dirichlet prior). The decomposable prior over structure and parameters is what [Dawid and Lauritzen, 1993] calls a *compatible family of hyper-Markov priors*, one for each model structure. In the approach of [Dawid and Lauritzen, 1993] as well as in HGC and [Meilă and Jordan, 2000] conjugate priors are used to enable one to *compare* between different structures. In contrast, our goal is to instead *average* under the prior.

We have replaced a fourth assumption made by HGC, namely the

requirement that every graph structure is possibile, with the weaker assumption 4 that G^{sup} is connected. Even this weaker form is not essential for our results. In section 7.1 we shall give a general formulation of the above theorem that dispenses altogether with the connectivity assumption.

To summarize, starting with the assumptions 1–3 and aiming mainly at obtaining a tractable and consistent prior representation, we have arrived at the conclusion that the prior has to be a product of Dirichlet distributions. This demonstrates that our initial requirement is essentially a drastic one; the restrictions on the prior should be understood as restrictions on the type of prior information about the model we are allowed to have. In the remainder of this section, we discuss the kind of restriction implied by assumptions 1–3. But before embarking onto this, let us note that from the computational perspective, the advantage is enormous: first, a decomposable prior is defined by order n^2 hyper-parameters only; second, the decomposable prior is a conjugate prior and its hyper-parameters can be updated efficiently; third, and most important, computing the normalization constant of a decomposable prior is tractable. So are related operations like averaging under the prior. The next section will present the latter issues in detail.

The parameter means under a Dirichlet distribution are [DeGroot, 1975]

$$< \theta_j >_{D(.|N'_j, j=1,\ldots r)} = \frac{N'_j}{\sum_j N'_j} \tag{36}$$

Hence a Dirichlet prior expresses knowledge about the values of the parameters' means with a certain "confidence" $N' = \sum_j N'_j$ which is the same for all parameters. For instance, the Dirichlet prior introduced here will be inadequate if we have two sources of prior knowledge, involving disjoint subsets of parameters, and having different associated confidences.

The parameter independence assumption corresponds to prior knowl-

edge equivalent to having seen only complete observations of x. Observations with missing data introduce dependencies among the parameters and thus violate this assumption.

HGC also point out that the likelihood equivalence assumption characterizes knowledge equivalent to having seen only data obtained by passive observation. If one's prior knowledge is obtained for instance by experiments, or from a study with randomized subjects, then this assumption may not hold. For a detailed explanation of this phenomenon, the reader is asked to consult HGC.

6 Bayesian learning with decomposable priors

6.1 Computing the posterior

Here we address the problem formulated in section 3, of computing the exact form of the posterior $P(T|\mathcal{D})$ given by (11) and reproduced here

$$P(T|\mathcal{D}) \;=\; \frac{P_0(T)\prod_{t=1}^{N} T(x^t)}{P(\mathcal{D})}$$

From equations (1) and (2) we know that the likelihood can be written as a product over tree edges. Theorem 4 proves the same about the decomposable prior. It follows then that the posterior $P(T|\mathcal{D})$ in equation (11) can also be factored over the edges of T. We shall see that in addition $P(T|\mathcal{D})$ is decomposable and the normalization constant $P(\mathcal{D}) = Z_{\mathcal{D}}$ can be computed tractably.

We shall use the following important property of a Dirichlet distribution: Assume a discrete variable z that takes values $1 \ldots r$ with probabilities $\theta = (\theta_1, \ldots \theta_r)$, a prior for θ that is Dirichlet with hyperparameters $N'(1), \ldots N'(r)$ and a set $\mathcal{D}_z$ of N independent observations for z, such that the value j appears $N(j)$ times in $\mathcal{D}_z$. Then,

the posterior of the parameters θ is (see e.g. [DeGroot, 1975]) is also a Dirichlet distribution with hyper-parameters $N'(j) + N(j)$.

$$P(\theta|\mathcal{D}_z) \;=\; D(\theta\,;\, N'(.) + N(.)) \tag{37}$$

This result applies immediately to the posterior of a tree. Let us denote by $N_{uv}(ij)$ and $N_v(j)$ the *sufficient statistics* of the sample $\mathcal{D}$, i.e. the number of times $u = i, v = j$ and respectively $v = j$ in $\mathcal{D}$. For the sake of simplicity, we denote the posterior counts with a double ', i.e

$$\begin{aligned} N'' &= N' + N \\ N''_v(j) &= N'_v(j) + N_v(j) \\ N''_{uv}(ij) &= N'_{uv}(ij) + N_{uv}(ij) \end{aligned}$$

Then, from (37) and theorem 4 we obtain

$$P(T|\mathcal{D}) = \frac{1}{Z_{\mathcal{D}} Z_\beta} \left(\prod_{uv \in E} \beta_{uv} \right) \prod_{v \in V} \prod_{i=1}^{r_{\mathrm{pa}(v)}} Z_v(i) D(\theta_{v|\mathrm{pa}(v)}(.|i); N''_{v,\mathrm{pa}(v)}(i\,.)) \tag{38}$$

The constants $Z_v(i),\ v \in V,\ i = 1, \ldots r_{\mathrm{pa}(v)}$ represent the ratios of Gamma functions below

$$Z_v(i) \;=\; \frac{\Gamma(N'_{\mathrm{pa}(v)}(i)) \prod_{j=1}^{r_v} \Gamma(N''_{v,\mathrm{pa}(v)}(ji))}{\Gamma(N''_{\mathrm{pa}(v)}(i)) \prod_{j=1}^{r_v} \Gamma(N'_{v,\mathrm{pa}(v)}(ji))}$$

and $Z_\beta = |Q(\beta)|$ is the normalization constant of the structure prior defined by equations (13) and (14). Hence, $P(T|\mathcal{D})$ is also decomposable, and its hyper-parameters are available directly from the hyper-parameters of the prior and the sufficient statistics of the sample.

It remains to find the value of the normalization constant $Z_{\mathcal{D}}$. As a first step, we will keep the structure E fixed and integrate over the parameters $\theta_{\overline{E}}$ in some directed structure $\overline{E}$ obtained from E. Since the parameters' posterior in (38) is already in the form of a normalized distribution, which consequently integrates to 1, we obtain after

a simple calculation:

$$\int P(T|\mathcal{D})d\theta_{\overline{E}} = \frac{1}{Z_{\mathcal{D}}}\frac{1}{Z_\beta}\frac{\Gamma(N')}{\Gamma(N'')}\left[\prod_{v\in V}\prod_{j=1}^{r_v}\frac{\Gamma(N''_v(j))}{\Gamma(N'_v(j))}\right] \tag{39}$$
$$\cdot \prod_{uv\in E}\left[\beta_{uv}\prod_{i=1}^{r_u}\frac{\Gamma(N'_u(i))}{\Gamma(N''_u(i))}\prod_{j=1}^{r_v}\frac{\Gamma(N'_v(j))}{\Gamma(N''_v(j))}\prod_{i=1}^{r_u}\prod_{j=1}^{r_v}\frac{\Gamma(N''_{uv}(ij))}{\Gamma(N'_{uv}(ij))}\right]$$

This quantity represents the marginal posterior $P(E|\mathcal{D})$; as required by likelihood equivalence, this is the same no matter how $\overline{E}$ is obtained from E. Note also that $P(E|\mathcal{D})$ decomposes into a product over the edges in E preceded by factors independent of E. We define the edge weights W_{uv} and the node weights W_v as

$$W_v = \prod_{j=1}^{r_v}\frac{\Gamma(N''_v(j))}{\Gamma(N'_v(j))}$$
$$W_{uv} = \frac{1}{W_u W_v}\prod_{i=1}^{r_u}\prod_{j=1}^{r_v}\frac{\Gamma(N''_{uv}(ij))}{\Gamma(N'_{uv}(ij))}$$

With these definitions, equation (39) can be written as

$$P(E\,|\,\mathcal{D}) = \frac{1}{Z_{\mathcal{D}}}\frac{1}{Z_\beta}\frac{\Gamma(N')}{\Gamma(N'')}\prod_{v\in V}W_v\cdot\prod_{uv\in E}(\beta_{uv}W_{uv}) \tag{40}$$

Now, to get $Z_{\mathcal{D}}$, we sum over all structures by applying theorem 1 with weights βW.

$$1 = \sum_E P(E|\mathcal{D}) = \frac{1}{Z_{\mathcal{D}}}\frac{1}{Z_\beta}\frac{\Gamma(N')}{\Gamma(N'')}\prod_{v\in V}W_v\cdot|Q(\beta W)| \tag{41}$$

Therefore

$$Z_{\mathcal{D}} = \frac{|Q(\beta W)|}{|Q(\beta)|}\frac{\Gamma(N')}{\Gamma(N'')}\prod_{v\in V}\prod_{j=1}^{r_v}\frac{\Gamma(N''_v(j))}{\Gamma(N'_v(j))} \tag{42}$$

Replacing $Z_{\mathcal{D}}$ in (40) we get

$$P(E\,|\,\mathcal{D}) = \frac{1}{|Q(\beta W)|}\prod_{uv\in E}\beta_{uv}W_{uv} \tag{43}$$

which is the same as (13) with $\beta \rightarrow \beta W$. Now we have completely defined the posterior distribution $P(T|\mathcal{D})$. The posterior probability of any tree distribution T can be computed analytically based on equations (38) and (42) while (43) gives the posterior of any tree structure E. Note that the weights W_{uv} are never 0, so that the support graph of the posterior distribution coincides with the G^{sup} of the prior.

To compute the posterior representation from the data set we need $\mathcal{O}(n^2 r_{MAX}^2 N)$ operations to obtain the sufficient statistics, $\mathcal{O}(n^2 r_{MAX}^2)$ to evaluate the edge weights W_{uv} and an additional $\mathcal{O}(n^3)$ to evaluate the normalization constant $Z_{\mathcal{D}}$ for a total of $\mathcal{O}(n^2 r_{MAX}^2 N + n^3)$ operations. After obtaining these, computing the posterior of a tree by (38) is $\mathcal{O}(n r_{MAX})$ and computing the posterior of a tree structure by (43) is $\mathcal{O}(n)$.

6.2 Bayesian model averaging

To perform model averaging in computing the probability of a new data point x one has to evaluate

$$P(x|\mathcal{D}) \;=\; \sum_{E} \left[\int T(x;\theta_{\overline{E}}) P(\theta_{\overline{E}}|\mathcal{D}) d\theta_{\overline{E}} \right] P(E|\mathcal{D}) \qquad (44)$$

where $\overline{E}$ represents any orientation of the tree structure E. Just as before, we can first integrate the above expression over the parameters for a fixed E and then perform a summation over structures. The former step yields

$$\begin{aligned}
\int T(x;\theta_{\overline{E}}) P(\theta_{\overline{E}}|\mathcal{D}) d\theta_{\overline{E}} &= \int \prod_{v\in V} T_{v|\mathrm{pa}(v)}(x_v|x_{\mathrm{pa}(v)}) \prod_{v\in V} \prod_{i=1}^{r_{\mathrm{pa}(v)}} D(\theta_{v|\mathrm{pa}(v)}(.|i)\,;\, N''_{v\mathrm{pa}(v)}(.i)) \\
&= \prod_{v\in V} \frac{N''_{v,\mathrm{pa}(v)}(x_v x_{\mathrm{pa}(v)})}{N''_{\mathrm{pa}(v)}(x_{\mathrm{pa}(v)})} \\
&= \underbrace{\frac{1}{N''} \prod_{v\in V} N''_v(x_v)}_{w_0(x)} \cdot \prod_{uv\in E} \underbrace{\frac{N''_{uv}(x_v x_u)}{N''_u(x_u) N''_v(x_v)}}_{w_{uv}(x)}
\end{aligned}$$

Again, we note that the result includes a structure independent factor $w_0(x)$ and a product of factors corresponding to the tree edges $w_{uv}(x)$. The final result is invariant to the particular orientation $\overline{E}$ of E. Summing over tree structures is a mere exercise now; we have

$$P(x|\mathcal{D}) = \sum_E \frac{w_0(x)}{|Q(\beta W)|} \prod_{uv \in E} \beta_{uv} W_{uw} w_{uv}(x) \quad (45)$$

$$= \frac{w_0(x)|Q(\beta W w(x))|}{|Q(\beta W)|} \quad (46)$$

The averaging requires computing the edge weights $w(x)$ and evaluating a determinant, so that the total computation is $\mathcal{O}(n^3)$. This is a relatively large value compared to the $\mathcal{O}(n)$ operations necessary to compute the likelihood of x under the ML or MAP tree.

The result generalizes readily to several independent observations.

7 Extensions

7.1 Disconnected support graph

Here we generalize the previous results to the case when the support graph G^{sup} is disconnected. In other words, we discuss the case when the prior $P(E)$ enforces probabilistic independence between the edges and parameters in different connected components.

The intuition behind the following results stems from the fact that, for a disconnected support graph, the domain V is effectively partitioned into K subsets of variables V^k, $k = 1, \dots K$ each corresponding to a connected component of G^{sup}. We denote these components by $G^k = (V^k, E^k)$, $k = 1, \dots K$. We also introduce a notation similar to (15), to denote a set of values corresponding to pairs of variables in a subset U of V.

$$a_U = \{a_{uv},\, u \neq v,\, u, v \in U \subseteq V\}$$

As we shall see, the subsets V^k behave as separate and independent domains from all points of view.

Because there can be no edges between the subsets, it is easy to see that

$$\beta \;=\; \bigcup_{k=1}^{K} \beta_{V^k} \tag{47}$$

This holds in general for a function f defined over E^{sup}.

A disconnected graph will have no spanning tree (hence the value returned by the Matrix Tree theorem will be 0) but it will have *maximal trees*[5], i.e. trees having a maximal number of edges. A maximal tree is composed of spanning trees corresponding to each G^k (and has a total number of $n-K$ edges). Hence, the number of distinct maximal trees in G^{sup} is the product of the numbers of spanning trees in its connected components:

$$\#\text{ maximal trees}(G^{sup}) \;=\; \prod_{k=1}^{K} \#\text{ spanning trees}(G^k) \tag{48}$$

The above remarks allow us to prove a generalization of theorem 1.

Theorem 5 *Let $P(E)$ be a distribution over maximal tree structures defined by (13,14) with $\beta \geq 0$. Then the normalization constant Z is equal to*

$$Z \;=\; \prod_{k=1}^{K} |Q(\beta_{V^k})| \tag{49}$$

The proof of this theorem is an easy consequence of theorem 1 and of the previous remarks and therefore it is omitted.

Intuitively, one can imagine redefining $Q(\beta)$ as a block diagonal matrix of dimension $n-K$ consisting of blocks $Q(\beta_{V^k})$, $k = 1, \ldots K$. Then, one can rewrite (49) so as to obtain a form identical with the one in theorem 1

$$Z \;=\; |Q(\beta)| \tag{50}$$

If one defines $M(\beta_{V^k})$ to correspond to each $Q(\beta_{V^k})$ in a way similar to equation (19), then one has

$$\frac{\partial}{\partial \beta_{uv}} |Q(\beta_{V^k})| \;=\; M_{uv}(\beta_{V^k}) |Q(\beta_{V^k})| \quad \text{for } u, v \in V^k$$

Now assembling $M(\beta_{V^k})$ for $k = 1, \ldots K$ in a block diagonal matrix $M(\beta)$ with

$$M_{uv}(\beta) \;=\; \begin{cases} M_{uv}(\beta_{V^k}), & \text{for } u, v \in V^k \\ 0, & \text{otherwise} \end{cases} \tag{51}$$

one can formally recover equation (18):

$$\frac{\partial Z}{\partial \beta_{uv}} \;=\; M_{uv}(\beta) \prod_{k=1}^{K} |Q(\beta_{V^k})| \;=\; M_{uv}(\beta) |Q(\beta)|. \tag{52}$$

It is also worth making the following point: $M(\beta_{V^k})$ (or $M(\beta)$ in the case of a connected G^{sup}) are defined only when $|Q(\beta_{V^k})|$ is non-zero. However, the derivative of $|Q(\beta_{V^k})|$ exists in all cases and is defined in terms of the minor determinants of the elements of $Q(\beta_{V^k})$. The minor A^*_{uv} of a square matrix A is obtained by deleting row u and column v from A and computing the determinant of the remaining matrix. Assume that to obtain $Q(\beta_{V^k})$ from $\overline{Q}(\beta_{V^k})$ we delete row and column n_k. Then, for every $u, v \in V^k$ we have

$$\frac{\partial}{\partial \beta_{uv}} |Q(\beta_{V^k})| \;=\; \begin{cases} Q^*_{uu}(\beta_{V^k}) + Q^*_{vv}(\beta_{V^k}) - 2Q^*_{uv}(\beta_{V^k}), & \text{if } u, v \neq n_k \\ Q^*_{vv}(\beta_{V^k}), & \text{if } v \neq n_k,\ u = n_k \\ Q^*_{uu}(\beta_{V^k}), & \text{if } u \neq n_k,\ v = n_k \end{cases} \tag{53}$$

Similarly to (21), for the average of an additive function f we obtain

$$< f(E) >_P = \sum_{k=1}^{K} \sum_{u,v \in V^k, u<v} f_{uv} \beta_{uv} M_{uv}(\beta_{V^k}) \;=\; \sum_{u<v} f_{uv} \beta_{uv} M_{uv}(\beta) \tag{54}$$

Finally, if g is multiplicative, equation (24) becomes

$$< g(E) >_P = \frac{\prod_{k=1}^{K} |Q(\beta_{V^k} g_{V^k})|}{\prod_{k=1}^{K} |Q(\beta_{V^k})|} \;=\; \frac{|Q(\beta g)|}{|Q(\beta)|} \tag{55}$$

7.2 Trees with fewer than $n-1$ edges

In the present paper we focus on tree graphical models whose structure is a connected graph[6]. Now we briefly discuss a slightly more general case, in which the structure of the graphical model is a set of edges E containing no cycles. Such a graphical model is called by extension a tree graphical model. In this section, we will apply the term *general* when we talk about tree structures and graphical models with $|E| \leq n-1$ and *spanning* when $|E| = n-1$.

It is easy to see that equations (1) and (2) defining the tree distribution apply to general tree graphical models. The decomposable prior over structures and parameters can be extended to general trees in a straightforward manner. So are Assumptions 1-4. For more details, the reader should consult [Meilă and Jordan, 2000]. However, for general trees we do not know of a graph theoretical result analog to the Matrix Tree theorem. Therefore, exact Bayesian model averaging over the family of general tree distributions is not possible (to date, at least).

7.3 Ensembles of trees

In this section we consider a new probability model, called *ensembles of trees* that naturally extends the tree graphical model. To best describe this model, imagine that a tree distribution is defined in two steps: first a set of parameters θ and second the structure E. Because E is not known at the time when we choose θ, we need to specify a parameter set that is sufficiently large, so that for any E we can afterwards extract from θ the actual set of parameters θ_E. This can be done easily following the same idea that allowed us to define a decomposable prior

in section 5. We choose

$$\begin{aligned}\theta &= \{\theta_{uv}(ij),\ u,v \in V, i = 1,\dots r_u, j = 1,\dots r_v\} \\ &\quad \cup \{\theta_v(j),\ v \in V, j = 1,\dots r_v\} \end{aligned} \tag{56}$$

such that

$$\begin{aligned}\sum_{i=1}^{r_u} \theta_{uv}(ij) &= \theta_v(j)\ \forall\, u \in V \\ \sum_{j=1}^{r_v} \theta_v(j) &= 1\ \forall v \in V \end{aligned} \tag{57}$$

Now, changing the notation of equation (1) to emphasize the dependence on θ and E, we write the tree distribution as

$$T(x|\theta, E) \ = \ \prod_{uv \in E} \frac{\theta_{uv}(x_u, x_v)}{\theta_u(x_u)\theta_v(x_v)} \prod_{v \in V} \theta_v(x_v) \tag{58}$$

The ensemble of trees $R(x)$ is a weighted average of all the possible tree distributions sharing the same parameters θ. To ensure tractability, the weights $P(E)$ will represent a decomposable distribution over spanning tree structures as in (13).

$$R(x) \ = \ \sum_E P(E) T(x|\theta, E) \tag{59}$$

If we use the notations

$$\omega_{uv}(x) = \frac{\theta_{uv}(x_u, x_v)}{\theta_u(x_u)\theta_v(x_v)} \tag{60}$$

$$\omega_0(x) = \prod_{v \in V} \theta_v(x_v) \tag{61}$$

for the edge dependent and respectively edge independent factors in (58) then, by theorem 1, $R(x)$ has an alternative, tractable form

$$R(x) \ = \ \omega_0(x) \frac{|Q(\omega(x)\beta)|}{|Q(\beta)|}$$

The ensemble of trees can be seen as a mixture model whose components are the trees over V parametrized by θ. The weighted averaging corresponds then to the presence of a hidden variable z taking as

many values as there are structures, each with probability $Pr[z = E] = P(E)$. Therefore, the (generalized) EM algorithm [Dempster et al., 1977] can be considered as a possibility for learning the parameters. We shall not pursue this issue in detail, but we will mention the following: the E step of the algorithm is tractable and straightforward given equation (59); the M-step however cannot be performed exactly and it is not known if the expression to be maximized has a unique local maximum.

But if we assume a set of complete observations $\mathcal{D}$ as before, the likelihood of this data set, denoted by $R(\mathcal{D})$, can be optimized w.r.t. the parameters θ and β by gradient ascent. We shall denote by $M_{uv}(\beta)$ and $M_{uv}(\beta\omega(x^t))$ respectively the values in equation (19) that correspond to $Q(\beta)$ and $Q(\beta\omega(x^t))$. Using lemma 2 we obtain

$$\frac{\partial \log R(\mathcal{D})}{\partial \beta_{uv}} = \sum_{t=1}^{N} \frac{\omega_{uv}(x^t) M_{uv}(x^t)}{|Q(\beta\omega(x^t))|} - N M_{uv}(\beta) \tag{62}$$

$$\frac{\partial \log R(\mathcal{D})}{\partial \theta_{uv}(ij)} = \frac{\beta_{uv}}{\theta_u(i)\theta_v(j)} \sum_{t: x_u^t = i, x_v^t = j} M_{uv}(x^t) \tag{63}$$

$$\frac{\partial \log R(\mathcal{D})}{\partial \theta_v(j)} = \frac{1}{\theta_v(j)} \sum_{t: x_v^t = j} [1 - \sum_{v' \in V} \omega_{vv'}(x) M_{vv'}(x^t)] \tag{64}$$

Note that the parameters θ need to satisfy (57) and therefore we will need to perform a constrained maximization of $R(\mathcal{D})$ using e.g. Lagrange multipliers; this method will converge to a local optimum of the log-likelihood.

8 Discussion

This paper has focused on decomposable priors for tree distributions. A decomposable prior is expressed as a product of factors, each corresponding to an edge of the tree. The same edge contributes the same amount in every tree structure that includes it. This property

allows representing the prior of any tree distribution T by order n^2 hyper-parameters.

Decomposable priors have been considered before, typically in the context of general Bayes nets, of which tree distributions are a subset. However, for general Bayes nets, (1) the prior cannot be expressed with a tractable number of parameters except in special cases; and (2) the normalization constant of the prior cannot be computed. For the family of trees, taken separately, both of the above negatives become affirmatives.

Our paper's main contribution is to show (2), i.e that for a decomposable prior over $\mathcal{T}_V$ the normalization constant Z can be expressed analytically in closed form *and* computed tractably. This is something new and unique, since, to our knowledge, there has not been any other result of this kind in graphical models.

It is well known that, for Bayes nets where (a) each node has no more than k parents, (b) the variable ordering is given and (c) each structure has the same prior, computing Z and other averaging operations are $\mathcal{O}(n^{k+1})$. However, the total number of models in this family is $\mathcal{O}(n^{k+1})$ hence polynomial, while the family $\mathcal{T}_V$ is superexponential.

From this result it follows that other "Bayesian learning" operations are tractable, including: updating the normalization constant of the posterior, computing the marginal $P(E)$ of a particular structure, model averaging for the marginal probability $P(x)$ of the next observation. We also give analytic expressions for the derivatives of Z w.r.t the edge weights and for averages of additive and multiplicative functions under factored priors. They pave the way toward a much larger range of averaging operations than the few enumerated in this paper.

A second contribution is to examine the decomposable prior restricted to the family of trees. By this we wanted to ensure that one

can define the prior in a consistent and generally "well-behaved" way. We showed that any hyper-parameter set with $\beta \geq 0$ and $N'_{uv}(.,.)$ satisfying (33,34) represents a well-defined and consistent prior.

We also showed that the decomposable prior for trees can be derived from a set of standard assumptions and we proved the unicity of the Dirichlet prior in this case. It is remarkable that, for trees, these standard assumptions, which parallel those of HGC, are sufficient to ensure tractability. In fact, these assumptions are no stronger than the assumptions of *functional independence* implicit in the original Chow and Liu algorithm [Chow and Liu, 1968, Meilă-Predoviciu, 1999].

Is worth highlighting again that these assumptions are restrictive, in the sense of drastically limiting the type of prior knowledge that can be used efficiently in the Bayesian learning of trees. Prior knowledge that cannot be expressed as a factored prior is e.g. knowledge that two edges are more likely to appear simultaneously than separately in a tree structure, or knowledge that two edges have the same parameters. This problem is not specific to trees, but to Bayes nets in general. Therefore, a worthwhile area of future research is discovering tractable methods to deal with such type of knowledge in the case of tree structures or in the case of general Bayes nets.

One should also be cautioned that not all model averaging operations are tractable under the decomposable prior. For example, while computing the marginal of a complete observation $P(x)$ was shown to be tractable, computing the marginal of an arbitrary subset of variables $U \subseteq V$ is generally intractable. This is due to the fact that marginalizing out the variables in $V \setminus U$ amounts to another averaging operation. It is known [Meilă and Jordan, 2000] that the average of several tree graphical models is not a graphical model. Consequently, the efficient methods of conditioning and marginalizing in a tree (or

more general graphical model) do not apply to this situation.

Finally, we have also introduced ensembles of trees as a tractable extension to the tree model. Ensembles of trees can be learned in the ML framework. Exploring the properties of the new model and of the learning algorithm itself are areas of continuing research.

Acknowledgements

The authors are grateful to Jesus Cerquides for discovering an error in an earlier verion of this manuscript. Most results in this paper were presented in compressed and preliminary form in [Meilă and Jaakkola, 2000]. Theorem 1 was first proved by the authors in [Jaakkola et al., 2000]; later we discovered a similar result in [Harary, 1967]. This work was funded by ONR contract number N00014-98-C-0326 and by NSF KDI award number DMS–9873442.

Appendix

Proof of Theorem 1 First we state the matrix tree theorem, on which our proof is based. The theorem is due originally to Kirchhoff, who published it as part of his work on electrical circuits.

A multigraph $G = (V, E)$ is a graph where E may contain more than one edge between the same two vertices (i.e E is a multiset of unordered pairs from $V \times V$).

Theorem 6 (Matrix Tree Theorem) [West, 1996] *Let $G = (V, E)$ be a multigraph and denote by $a_{uv} = a_{vu}$ the number of undirected edges between vertices u and v. Then the number of all spanning trees of G is given by $|A_{uv}|(-1)^{u+v}$ the value of the determinant obtained*

from the following matrix by removing row u and column v.

$$A = \begin{bmatrix} \deg v_1 & -a_{12} & -a_{13} & \dots & -a_{1,n} \\ -a_{21} & \deg v_2 & -a_{23} & \dots & -a_{2,n} \\ \dots & \dots & \dots & \dots & \dots \\ -a_{n,1} & -a_{n,2} & -a_{n,3} & \dots & \deg v_n \end{bmatrix}$$

Our result is the generalization of the matrix tree theorem for a real-valued (and renamed to β). We shall prove it first for positive integer values of β, then for positive rational values. Then, because the determinant is a continuous function it will follow that the theorem is true for any real, positive values β.

Assume β are integers. Then β_{uv} can represent the number of edges between u and v of a multigraph over V. The degree of node v equals the number of edges incident to v; this number is $\sum_{u \neq v} \beta_{uv}$. Thus, by theorem 6, the total number of spanning trees in this graph is $Z_1 = |Q(\beta)|$. We now show that $Z_1 = Z$ (where Z is defined by (14)).

$$\begin{aligned} Z_1 &= \sum_E \#\, distinct\, trees\, having\, structure\, E \qquad (65) \\ &= \sum_E \prod_{uv \in E} \beta_{uv} \\ &= Z \end{aligned}$$

Assume now that β are all rational. Let m be the common denominator of β, such that $\beta_{uv} = a_{uv}/m$ with a_{uv} integer. Then $\overline{Q}(a) = m\overline{Q}(\beta)$ is a matrix of integers and by virtue of (65) we have that

$$|Q(a)| = \sum_E \prod_{uv \in E} a_{uv} \qquad (66)$$

But $|mQ(\beta)| = m^{n-1}|Q(\beta)|$ and therefore

$$\begin{aligned} |Q(\beta)| &= \frac{1}{m^{n-1}} \sum_E \prod_{uv \in E} a_{uv} \\ &= \sum_E \prod_{uv \in E} \frac{a_{uv}}{m} \\ &= \sum_E \prod_{uv \in E} \beta_{uv} \\ &= Z \end{aligned}$$

■

Proof of lemma 2 The proof uses the fact that, for any nonsingular matrix with elements A_{ij}

$$\frac{\partial |A|}{\partial A_{ij}} = |A|(A^{-1})_{ij} \tag{67}$$

Then, for $u, v < n$, taking into account that the only elements of $Q(\beta)$ that depend on β_{uv} are Q_{uu}, Q_{vv}, Q_{uv} and Q_{vu} we have successively:

$$\frac{\partial Z}{\partial \beta_{uv}} = \frac{\partial |Q(\beta)|}{\partial \beta_{uv}} = \sum_{i,j \in V} \frac{\partial |Q(\beta)|}{\partial Q_{ij}} \frac{\partial Q_{ij}}{\partial \beta_{uv}} = |Q(\beta)|[(Q^{-1})_{uu} + (Q^{-1})_{vv} - 2(Q^{-1})_{uv}] \tag{68}$$

Similarly, if $u = n$, $v < n$, then β_{vn} appears only in Q_{vv}. Hence

$$\frac{\partial Z}{\partial \beta_{uv}} = |Q(\beta)|(Q^{-1})_{vv} \tag{69}$$

■

Proof of lemma 3 We first introduce the following lemma:

Lemma 7 *If $P(E)$ is given by equation (13) and f is an additive function of E then*

$$< f(E) >_P = \frac{1}{Z} \left. \frac{\partial |Q(\beta e^{\alpha f})|}{\partial \alpha} \right|_{\alpha = 0} \tag{70}$$

This lemma can be easily proved by equating $Q(\beta e^{\alpha f})$ with its definition (16) and then taking derivatives of both sides.

Then, to obtain equation (22) we use (67) again, conveniently grouping the terms involving f_{uv} afterwards. To derive to compact form (23) we notice that (22) represents the sum of element-wise products of $Q(f\beta)$ and $Q^{-1}(\beta)$ and use the well-known matrix identity:

$$\sum_{ij} A_{ij} B_{ij} \;=\; \text{trace} AB^T$$

■

Proof of theorem 4 Begin by fixing two variables $u, v \in V$ such that $\beta_{uv} > 0$ and a structure E containing edge uv. Denote by $\overline{E}^1$ and $\overline{E}^2$ the orientations of E where u and respectively v are root. Thus $\overline{E}^1$ and $\overline{E}^2$ differ only in the orientation of edge uv. Let θ^1 and θ^2 be two parametrizations corresponding to $\overline{E}^1$ and $\overline{E}^2$ respectively, such that θ^1 and θ^2 produce the same distribution T. For every θ^1 there will be a unique θ^2 (obtained by applying (1) and (2)) satisfying this condition. Moreover,

$$\theta^1_{w|\text{pa}(w)} \;=\; \theta^2_{w|\text{pa}(w)} \quad \text{for all } w \neq u, v. \tag{71}$$

The prior distribution of θ^1 is

$$P_0^1(\theta^1) \;=\; P_0^1(\theta^1_u) P_0^1(\theta^1_{v|u}) \prod_{w \neq u,v} P_0^1(\theta^1_{w|\text{pa}(w)}) \tag{72}$$

Similarly, for θ^2 we have

$$P_0^2(\theta^2) \;=\; P_0^2(\theta^2_v) P_0^2(\theta^2_{u|v}) \prod_{w \neq u,v} P_0^2(\theta^2_{w|\text{pa}(w)}) \tag{73}$$

The likelihood equivalence assumption, together with the change of variable formula, imply

$$P_0^1(\theta^1) \;=\; P_0^2(\theta^2(\theta^1)) \left| \; \tfrac{\partial \theta^2}{\partial \theta^1} \; \right| \tag{74}$$

In the above, $|\frac{\partial\theta^2}{\partial\theta^1}|$ denotes the Jacobian of the transformation $\theta^1 \rightarrow \theta^2$. Due to the equality (71) this Jacobian has the following structure:

$$|\frac{\partial\theta^2}{\partial\theta^1}| = \begin{vmatrix} \frac{\partial\theta^2_v}{\partial\theta^1_u} & \frac{\partial\theta^2_{u|v}}{\partial\theta^1_u} & 0 & 0 & \dots & 0 \\ \frac{\partial\theta^2_v}{\partial\theta^1_{v|u}} & \frac{\partial\theta^2_{u|v}}{\partial\theta^1_{v|u}} & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{vmatrix} = \begin{vmatrix} \frac{\partial\theta^2_v}{\partial\theta^1_u} & \frac{\partial\theta^2_{u|v}}{\partial\theta^1_u} \\ \frac{\partial\theta^2_v}{\partial\theta^1_{v|u}} & \frac{\partial\theta^2_{u|v}}{\partial\theta^1_{v|u}} \end{vmatrix}$$

Let us denote the resulting determinant by $\left|\frac{\partial(\theta^2_v,\theta^2_{u|v})}{\partial(\theta^1_u,\theta^1_{v|u})}\right|$. The above equality together with (72–74) implies

$$P^1_0(\theta^1_u,\theta^1_{v|u}) \;=\; P^2_0(\theta^2_v(\theta^1_u,\theta^1_{v|u}),\theta^2_{u|v}(\theta^1_u,\theta^1_{v|u}))\left|\;\frac{\partial(\theta^2_v,\theta^2_{u|v})}{\partial(\theta^1_u,\theta^1_{v|u})}\;\right| \qquad (75)$$

Now we have reached our first partial goal, because by theorem 7 in HGC equation (75) implies that $P_0(\theta_{uv})$ as well as $P^1_0(\theta^1_u,\theta^1_{v|u})$ and $P^2_0(\theta^2_v,\theta^2_{u|v})$ are Dirichlet. Let us denote by $N'_{uv}(ij)$, $i = 1,\dots r_u$, $j = 1,\dots r_v$ the hyper-parameters of $P(\theta_{uv})$ and by $N'_u(i)$ and $N'_v(j)$ their sums over j and i respectively. It is now easy to show that

$$P^1_0(\theta^1_u) \;=\; D(\theta^1_u(.)\,;\,N'_u(.)) \qquad (76)$$

$$P^1_0(\theta^1_v|u) \;=\; \prod_{i=1}^{r_u} D(\theta^1_{v|u}(.|i)\,;\,N'_{uv}(i.)) \qquad (77)$$

$$P^2_0(\theta^2_v) \;=\; D(\theta^2_v(.)\,;\,N'_v(.)) \qquad (78)$$

$$P^2_0(\theta^2_u|v) \;=\; \prod_{j=1}^{r_v} D(\theta^2_{u|v}(.|j)\,;\,N'_{uv}(.j)) \qquad (79)$$

By parameter modularity, these identities are true for any tree structure containing an edge between u and v. Therefore, all that remains to be shown is that $N' = \sum_{j\in\Omega_v} N'_v(j)$ has the same value for all $v \in V$.

Let $u \neq v$ be variables in V. We shall prove that

$$\sum_{i \in \Omega_u} N'_u(i) = \sum_{j \in \Omega_v} N'_v(j) \tag{80}$$

We distinguish two cases: $\beta_{uv} > 0$ and $\beta_{uv} = 0$. The former is already solved, by virtue of the first part of the proof. It remains to show that (80) holds when $\beta_{uv} = 0$ and no tree structure will contain edge uv.

Indeed, because G^{sup} is connected, there will exist at least a path between u and v in G^{sup}. Now we can easily show (80) by induction over the length of the path. ■

This theorem and its proof are easily extended to multiply connected G^{sup}. In that case, each component of the graph will have its own equivalent sample size N'^k.

References

[Cheeseman and Stutz, 1995] Cheeseman, P. and Stutz, J. (1995). Bayesian classification (AutoClass): Theory and results. In Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180. AAAI Press.

[Chow and Liu, 1968] Chow, C. K. and Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467.

[Cooper and Herskovits, 1992] Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

[Dasgupta, 1999] Dasgupta, S. (1999). Learning polytrees. In Laskey, K. B. and Prade, H., editors, *Proceedings of the 15th Conference on Uncertainty in AI*, San Francisco, CA. Morgan Kaufmann.

[Dawid and Lauritzen, 1993] Dawid, A. P. and Lauritzen, S. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 21:1272–1317.

[DeGroot, 1975] DeGroot, M. H. (1975). *Probability and Statistics*. Addison–Wesley Pub. Co., Reading, MA.

[Dellaportas and Forster, 1999] Dellaportas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86(3):615–633.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38.

[Giudici and Green, 1999] Giudici, P. and Green, P. (1999). Decomposable graphical gaussian model determination. *Biometrika*, 86:785–801.

[Harary, 1967] Harary, F. (1967). Graphs and matrices. *SIAM Review*, 9(1):83–90.

[Heckerman et al., 1995] Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.

[Jaakkola et al., 2000] Jaakkola, T., Meilă, M., and Jebara, T. (2000). Maximum entropy discrimination. In Solla, S. A., Leen, T. K., and Müller, K.-R., editors, *Neural Information Processing Systems*, volume 12, pages 470–476. MIT Press.

[Madigan and Raftery, 1994] Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89:1335–1346.

[Meilă and Jaakkola, 2000] Meilă, M. and Jaakkola, T. (2000). Tractable bayesian learning of tree distributions. In Boutilier, C. and Goldszmidt, M., editors, *Proceedings of the 16th Conference on Uncertainty in AI*, pages 380–388, San Francisco, CA. Morgan Kaufmann.

[Meilă and Jordan, 2000] Meilă, M. and Jordan, M. I. (2000). Learning with mixtures of trees. *Journal for Machine Learning Research*, 1(1).

[Meilă-Predoviciu, 1999] Meilă-Predoviciu, M. (1999). *Learning with Mixtures of Trees*. PhD thesis, Massachusetts Institute of Technology.

[Murray and Rice, 1993] Murray, M. K. and Rice, J. W. (1993). *Differential geometry and statistics*. Chapman & Hall.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman Publishers, San Mateo, CA.

[Spiegelhalter and Lauritzen, 1990] Spiegelhalter, D. J. and Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:491–505.

[Spirtes and Meek, 1995] Spirtes, P. and Meek, C. (1995). Learning Bayesian networks with discrete variables from data. In *Proceedings of First International Conference on Knowledge Discovery and Data Mining,* Montreal, QU, pages 294–299. Morgan Kaufmann.

[Srebro, 2001] Srebro, N. (2001). Maximum likelihood bounded tree-width Markov networks. In Breese, J. and Koller, D., editors, *Proceedings of the 17th Conference on Uncertainty in AI*, pages 504–511, San Francisco, CA. Morgan Kaufmann.

[West, 1996] West, D. B. (1996). *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ.

Notes

[1]University of Washington, Department of Statistics, mmp@stat.washington.edu

[2]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, tommi@csail.mit.edu

[3]This means that the size of the family, as a function of the number of variables n, grows faster than any polynomial in n.

[4]Each undirected E is one equivalence class.

[5]Maximal trees are called *maximal forests* in [West, 1996].

[6]We assume that G^{sup} is connected. The generalization to a disconnected G^{sup} is immediate, cf section 7.1.