Multiplicative Factorization of Multi-Valued NIN-AND Tree Models

Yang Xiang and Yiting Jin
University of Guelph, Canada

Abstract

A multi-valued Non-Impeding Noisy-AND (NIN-AND) tree model has the linear complexity and is more expressive than common Causal Independence Models (CIMs). We formulate a Multiplicative Factorization (MF) for multi-valued NIN-AND Tree (NAT) models. In comparison with the MF for binary NAT models (of a undirected tree structure), the proposed MF is a hybrid and multiply connected graphical model. Although a NAT is made of two types of NIN-AND gates, we show that a sound and space efficient MF requires multiple types of gate MFs, and therefore significantly more sophisticated parameterization and integration of gate MFs, and soundness analysis. We show that the formulated MF is exact and its space complexity is linear on the number n of causes per effect. Based on the proposed MF, we extend the scheme for lazy propagation (LP) with binary NAT-modeled Bayesian Networks (BNs) to multi-valued NAT-modeled BNs. We show that the extended scheme is more powerful than LP based on MF of noisy-MAX. We demonstrate that the scheme allows significantly more efficient LP both in space and in time.

1 Introduction

A BN quantifies the causal strength between an effect and its n causes by a Conditional Probability Table (CPT) whose number of parameters is exponential in n. Common CIMs, e.g., noisy-OR (Pearl 1988), reduce the number to being linear in n, but are limited in expressiveness. NAT models (Xiang and Jia 2007) have a linear number of parameters, and express both reinforcement and undermining as well as their recursive mixture. CIMs are not directly operable by common BN inference algorithms, e.g., the cluster tree method (Jensen, Lauritzen, and Olesen 1990). A number of techniques have been proposed to overcome the difficulty, e.g., (Zhang and Poole 1996; Madsen and D'Ambrosio 2000). One technique is MF (Takikawa and D'Ambrosio 1999) and related tensor decomposition (Savicky and Vomlel 2007).

To take advantage of NAT models in inference, MF has been applied to binary NAT models (Xiang 2012a). However, binary NAT models are not sufficiently general. Advancing MF from binary to multi-valued NAT models (Xiang 2012b) encounters several issues. The number of auxiliary variables per NIN-AND gate increases from one in the binary case to multiple. The dependency structure per gate changes from a undirected star to a multiply connected hybrid graph. The MF of a NAT is integrated from gate MFs as in the binary case, but alternative types of gate MFs increase from two to multiple. Parameterization over the hybrid structure with sound and space efficient coordination between multiple types of gate MFs requires more sophisticated design and analysis. We present our solution to these issues and a general MF for multi-valued NAT models. We show its soundness and demonstrate significantly more efficient inference in time and space when the MF is applied to multi-valued NAT-modeled BNs.

Sec. 2 covers background on NAT models. Sec. 3 outlines the technical challenges. Sec. 4 shows that NAT models are more general than noisy-MAX. The MF of NAT models is developed in Secs. 5 to 8. Its soundness and space complexity are analyzed in Sec. 9. Compilation of binary NAT-modeled BNs is extended to multi-valued NAT-modeled BNs for LP in Sec. 10. Experimental evaluation is described in Sec. 11. We omit proofs for space.

2 Background

Consider an effect e and the set of all causes $C = \{c_1, ..., c_n\}$, all of which are multi-valued and graded. That is, e has a domain $D_e = \{e^0, ..., e^\eta\}$ $(\eta \geq 1)$, where a higher index signifies a higher intensity, e^0 is *inactive*, and $e^1, ..., e^\eta$ are *active*. The domain of c_i is $D_i = \{c_i^0, ..., c_i^m\}$.

We categorize a causal event as *success* or *failure* depending on whether e is rendered active at certain intensity, as *single-causal* or *multi-causal* depending on the number of active causes, and as *simple* or *congregate* depending on the range of effect values. For instance, $P(e^k \leftarrow c_i^j)$ $= P(e^k|c_i^j, c_z^0 : \forall z \neq i)$ $(j > 0)$ is the probability of a *simple single-causal success*. $P(e \geq e^k \leftarrow c_1^{j_1}, ..., c_q^{j_q})$ $= P(e \geq e^k|c_1^{j_1}, ..., c_q^{j_q}, c_z^0 : c_z \in C \setminus X)$ (each $j > 0$) is the probability of a *congregate multi-causal success*, where $X = \{c_1, ..., c_q\}$ $(q > 1)$. It is also denoted by $P(e \geq e^k \leftarrow \underline{x}^+)$. Probability of the *null causal event* is $P(e^k \leftarrow \perp) = P(e^k|c_i^0 : \forall i) = 1$ for $k = 0$ and 0 for $k > 0$.

There are two types of multi-valued NIN-AND gates, each of which involves disjoint sets of causes $W_1, ..., W_q$. An input event of a *direct* gate is $e \geq e^k \leftarrow \underline{w}_i^+$ and the output event is $e \geq e^k \leftarrow \underline{w}_1^+, ..., \underline{w}_q^+$. An input of a *dual* gate is $e < e^k \leftarrow \underline{w}_i^+$ and the output event is $e < e^k \leftarrow \underline{w}_1^+, ..., \underline{w}_q^+$. Probability of the output event of a gate is the product of probabilities of its input events.

Causal interactions can be characterized as reinforcing or undermining based on the magnitude of causal probability of a set of active causes relative to those of its proper subsets. A direct gate models undermining causal interactions, and a dual gate models reinforcing. A multi-valued NAT organizes multiple gates into a tree to express mixture of reinforcing and undermining recursively. More details on multi-valued NAT models can be found in (Xiang 2012b).

3 Technical Issues

To take advantage of NAT models in inference, MF of binary NAT models is developed (Xiang 2012a). For each type (direct and dual) of gates, a gate MF is made of a undirected star structure with a single auxiliary variable and with one potential assigned to each link. MF of a binary NAT integrates the two types of gate MFs. Its structure is a undirected tree.

Construction of MF for multi-valued NAT Models face a number of issues. What is the dependence structure of a (stand-alone) multi-valued NIN-AND gate? As shown below, a suitable structure is multiply connected and is a hybrid graph with multiple auxiliary variables. How should each type of gate be parameterized? It turns out that although parameterization of a dual gate is probabilistic (based solely on causal probabilities), that of a direct gate must be pseudo-probabilistic (see below). To allow each type of gate MF to interface with the other in a NAT, variable domains and parameters of the gate MF must be redefined. A gate may be the leaf gate in a NAT or feeds into another. To render gate MFs space efficient, a distinct gate MF is needed in each case. As a result, four types of gate MFs are needed in the MF of a NAT model. Out of the $C(4, 2) = 6$ possible interactions among these gate MFs, which ones are valid and where they should apply? How should each type of gate MF and each type of valid interface be parameterized to allow sound coordination? Finally, a sophisticated analysis is necessary to establish soundness of the MF resultant from interactions of all gate MFs in a NAT model. After relating the dual gates to noisy-MAX in the next section, we present solutions to the above issues.

4 Equivalence of Dual Gate and Noisy-MAX

We show that NAT models generalize noisy-MAX (Henrion 1989; Diez 1993). In particular, noisy-MAX models are equivalent to multi-valued dual NIN-AND gate models, and hence are a special case of multi-valued NAT models.

Theorem 1 *Let $X = \{c_1, ..., c_n\}$ $(n \geq 1)$ be a set of causes of effect e that interact according to noisy-MAX. Let g be a dual NIN-AND gate where each input event involves exactly one c_i $(i = 1, ..., n)$. Then the causal probability of the output event of g is identical to that of noisy-MAX.*

By Theorem 1, a noisy-MAX can always be expressed as a dual NIN-AND gate. Since a dual gate models reinforcing, so does a noisy-MAX. Since NATs model reinforcing and undermining as well as their recursive mixtures, they are strictly more expressive than noisy-MAX models.

5 MF of Dual Gate Models

We organize MF of a dual gate model according to a hybrid graph G (Fig. 1), whose nodes are labeled by the effect $e \in \{e^0, e^1, ..., e^\eta\}$, causes $c_i \in \{c_i^0, ..., c_i^m\}$ $(i = 1, ..., n)$, and auxilary variables d_j $(j = 1, ..., \eta)$ (one for each active value of e), where $d_j \in \{0, 1\}$. The link between each pair of c_i and d_j is undirected, called a *clink*, as c_i is a cause variable. The link between each d_j and e is directed. The link type determines how potentials are defined as follows.

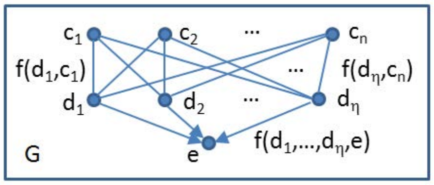

Figure 1: Hybrid graphical model for MF of a gate.

Each clink is assigned a potential $f(d_j, c_i)$. Node e is assigned a general potential (with negative values) $f(d_1, ..., d_\eta, e)$ over its family defined by incoming directed links, called the *family potential*. Table 1 specifies the MF potentials. We refer to the collection of graph G and the potentials as the MF of a Dual gate model (MDu).

Table 1: The clink and family potentials of MDu

d_j	c_i	f
0	c_i^0	1
0	c_i^1	$P(e < e^j \leftarrow c_i^1)$
	...	...
0	c_i^m	$P(e < e^j \leftarrow c_i^m)$
1	c_i	1

$(d_1, ..., d_\eta, e)$	f
$d_i = 0,\ \forall_{j \neq i}\ d_j = 1,$ $e = e^{i-1}$	1
$d_i = 0,\ \forall_{j \neq i}\ d_j = 1,$ $e = e^i$	-1
$\forall_i\ d_i = 1,\ e = e^\eta$	1
otherwise	0

We often obtain a product of potentials and then may marginalize out some variables, such as

$$f(e, c_1, ..., c_n) = \sum_{d_1,\ldots,d_\eta} f(d_1, ..., d_\eta, e) \prod_{1 \leq j \leq \eta, 1 \leq i \leq n} f(d_j, c_i).$$

We refer to the result as a Marginalized Potential Product (MPP). When relevant potentials are clear, we mention the MPP, e.g., $f(e, c_1, ..., c_n)$, without listing the potentials.

By Theorem 1, MDu is equivalent to MF of noisy-MAX from which the soundness of MDu follows.

Corollary 1 *Let MDu be applied to a dual NIN-AND gate model whose CPT is $P(e|c_1, ..., c_n)$. The MPP from the MDu satisfies $f(e, c_1, ..., c_n) = P(e|c_1, ..., c_n)$.*

Although MDu is equivalent to MF of noisy-MAX, it differs from previous work. The MF in (Takikawa and D'Ambrosio 1999) is not a graphical model (potentials are not structured through graphs). The MF in (Madsen and D'Ambrosio 2000) uses a DAG but its potential assignment does not follow family convention (each child variable is assigned a potential over itself and its parents). In fact, neither DAGs nor undirected graphs can suitably express the dependency. MDu is a hybrid graphical model with a rigorous syntax, where a potential is assigned to each clink and to the family when links are directed.

6 MF of Direct Gate Models

MF of a direct gate model is also structured as G in Fig. 1, but each $d_j \in \{0, 1, 2\}$ is ternary (see below). Table 2 specifies MF potentials. The collection of graph G and the potentials is referred to as the MF of a Direct gate model (MDi).

Table 2: The clink and family potentials of MDi

d_j	c_i	f
0	c_i^0	1
0	c_i^1	$P(e \geq e^j \leftarrow c_i^1)$
	...	...
0	c_i^m	$P(e \geq e^j \leftarrow c_i^m)$
1	c_i	1

d_j	c_i	f
2	c_i^0	1
2	c_i^1	0
	...	...
2	c_i^m	0

line	$(d_1, ..., d_\eta, e)$	f
1	$d_i = 0,\ \forall_{j \neq i}\ d_j = 1,\ e = e^{i-1}$	-1
2	$d_i = 0,\ \forall_{j \neq i}\ d_j = 1,\ e = e^i$	1
3	$\forall_i\ d_i = 1,\ e = e^0$	1
4	$d_1 = 2,\ \forall_{i>1}\ d_i = 0,\ e = e^0$	1
5	$d_1 = 2,\ \forall_{i>1}\ d_i = 0,\ e = e^\eta$	-1
6	otherwise	0

Note $f(d_j = 0, c_i = c_i^0) = 1 \neq P(e \geq e^j \leftarrow c_i^0) = 0$. That is, unlike in MDu, $f(d_j = 0, c_i)$ is not made entirely of probabilities. We refer to MDu as being probabilistic and MDi as being *pseudo-probabilistic*. It is necessary as value 0 blocks other potential values in computing MPP. To ensure soundness, the domain size of d_j is also larger in MDi to allow necessary manipulation (see $f(d_j = 2, c_i)$ and lines 4 and 5 in Table 2). Theorem 2 states its soundness.

Theorem 2 *Given a direct NIN-AND gate model with CPT $P(e|c_1, ..., c_n)$, the MPP from its MDi satisfies $f(e, c_1, ..., c_n) = P(e|c_1, ..., c_n)$.*

7 MF of NIN-AND Tree Models

A nontrivial NAT has at least two gates, e.g., Fig. 2 (a), where event labels are simplified and ovals into gates are omitted. MF of a NAT model consists of a hybrid graph G and a set of potentials defined over each undirected link and each family in G. G is integrated from graphs of gate MFs according to NAT topology, as shown in (b). Gate g_4 in (a) induces the subgraph spanning $\{c_1, c_2, a_1, a_2, b_1\}$ in (b). The child variable from the MF of the leaf gate is the *effect* variable and labeled by e, e.g., the leaf gate g_1 in (a). Child variables from MFs of other gates are *internal* variables and labeled differently. For instance, the child variable of MF for g_2 in (a) is labeled as b_3 in (b). The structure G of the MF for a multi-valued NAT differs significantly from that for a binary NAT (Xiang 2012a), in that the latter is a undirected tree while the former is hybrid and multiply connected.

The subgraph of G induced by a gate is identical to the MF graph for a standalone gate. However, variable domains and potentials associated with the subgraph may differ from those in the MF of a standalone gate for several reasons. First, an undirected link, e.g., $\langle k_1, b_3 \rangle$ in Fig. 2 (b), may connect an internal variable b_3. It is thus called an *ilink* and its potential must differ from that of a clink.

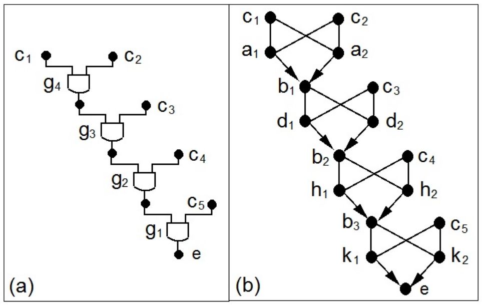

Figure 2: (a) A 4-gate NAT. (b) MF graph of (a).

Second, gates have different levels. The leaf gate is at level 1, e.g., g_1. A gate feeding the leaf gate is at level 2, e.g., g_2. The MF of a gate must adapt to its level. For instance, MF of a dual gate at level 2 is more sophisticated than that of a leaf gate. This is because the latter is terminal while the former feeds into a gate at the next level.

Third, all gates at the same level have the same type (dual or direct) and gates at adjacent levels differ in types. Hence, a dual gate at level 2 receives input from direct gates at level 3, and feeds into the direct leaf gate at level 1. The MF of a gate must be specified according to the gate from which it receives input and the gate that it feeds into.

Fourth, a gate can receive input from both clinks and ilinks. For instance, the MF of g_3 in Fig. 2 (b) receives input from clinks $\langle c_3, d_i \rangle$ $(i = 1, 2)$ as well as from ilinks $\langle b_1, d_i \rangle$. For the family potential over $\{d_1, d_2, b_2\}$ to work with both types of input uniformly, the product of potentials over $\langle c_i, a_j \rangle$ $(i, j = 1, 2)$, $\{a_1, a_2, b_1\}$, and $\langle b_1, d_i \rangle$, after marginalizing out $\{a_1, a_2, b_1\}$, must be equivalent (syntactically and semantically) to the product of potentials over $\langle c_3, d_i \rangle$. Suppose that gate g_3 in Fig. 2 (a) is direct and g_4 is dual. Recall that a clink potential of a direct gate is pseudo-probabilistic. To render the MPP probabilistic (Theorem 2), the domain size of auxiliary variables is increased from 2 in MDu to 3 in MDi. Here, output from MF of the dual gate g_4 is probabilistic (Corollary 1). To be equivalent to clink potentials over $\langle c_3, d_i \rangle$, the output needs to be rendered pseudo-probabilistic. As a result, the output variable b_1 of g_4 needs to have a domain larger than e, which will affect both the family potential for the MF of gate g_4 and the ilink potentials for the MF of gate g_3.

Due to these factors, four types of gate MFs are developed: MDu enhanced with ilink potentials, MDi enhanced with ilink potentials, Extended MDu (EDu), and Extended MDi (EDi). A particular gate is assigned one of them depending on its gate type (dual or direct) and level. The rule of assignment is illustrated in Table 3.

[Level 1] If a gate is the leaf gate, its MF is MDu or MDi depending on the type of gate. In either case, the output vari-

Table 3: Rule of MF assignment for NIN-AND gates

level of gate	type of gate MF	
3 or higher	EDu	EDi
2	EDu	MDi
1	MDu	MDi
	dual	direct

able is e, and the MPP output is probabilistic.

[Level 2] MDi requires pseudo-probabilistic input. For a dual gate at level 2 to feed into MDi at level 1, it must output accordingly. Its MF is EDu. If a direct gate is at level 2, it takes pseudo-probabilistic input and delivers output probabilistically. Hence, its MF is MDi.

[Level 3+] A dual gate at level 3 or higher plays the same role as at level 2. Hence, its MF is also EDu. EDi is described below as we present MF variables.

For example, if g_1 of Fig. 2 (a) is direct, the MF assignment for gates is (g_1: MDi; g_2: EDu; g_3: EDi; g_4: EDu). If g_1 is dual, it is (g_1: MDu; g_2: MDi; g_3: EDu; g_4: EDi).

Each gate MF has 4 types of variables. Denote an input cause by c, an input internal variable by b, an auxiliary variable by d, and the output (child) variable by h. Their domain sizes depend on the gate MF and is summarized in Table 4.

Table 4: Domain sizes for variables b, d and h in a gate MF

	domain size			
var	MDu	EDu	MDi	EDi
b	$\eta+1$	$\eta+2$	$\eta+2$	$\eta+2$
d	2	3	3	3
h		$\eta+2$	$\eta+1$	$\eta+2$

MDu takes probabilistic input from MDi and outputs probabilistically. Hence, d_{MDu} is binary, where subscript indexes the MF, and the domain size for b_{MDu} and h_{MDi} is the same as e. MDi takes pseudo-probabilistic input but outputs probabilistically. Hence, d_{MDi} is ternary, and the domain size for h_{EDu} and b_{MDi} is larger than e (by one).

Edu takes probabilistic input from a direct gate and outputs pseudo-probabilistically. Hence, d_{EDu} is ternary. As its output property differs from MDu, its clink potentials (see below) also differ from those of MDu. For the MPP output from an incoming direct gate to match syntax and semantics of clink potentials, b_{EDu} also differs from b_{MDu} in domain size (larger by one).

Since EDi feeds into EDu, h_{EDi} must match b_{EDu} in domain size and thus differ from h_{MDi}. Hence, EDi must be a distinct MF from MDi. EDi takes pseudo-probabilistic input from EDu at a higher level, and outputs probabilistically to EDu at a lower level. Hence, d_{EDi} is ternary, and b_{EDi} has the same domain size as h_{EDu}.

As presented above, domain setup for each type of gate MF has aimed at keeping the domain size of each MF variable as small as possible. Since each type of gate MF is repeatedly applied in NAT modeling, this effort will pay off in the overall space complexity of MF for NAT models.

8 Potentials in MF of NAT Models

Potentials for each type of gate MF are specified below.

[**MDu**] The clink and family potentials of MDu are shown in Table 1. The ilink potential of MDu is shown in Table 5.

Table 5: The ilink potential $f(d_j, b)$ $(j = 1, ..., \eta)$ of MDu

line	(d_j, b)	f
1	$d_j = 0,\ b = b^0, ..., b^{j-1}$	1
2	$d_j = 0,\ b = b^j, ..., b^\eta$	0
3	$d_j = 1$	1

[**EDu**] The clink, ilink and family potentials of EDu are shown in Table 6.

Table 6: The clink, ilink and family potentials $f(d_j, c_i)$, $f(d_j, b)$ and $f(d_1, ..., d_\eta, h)$ of EDu

d_j	c_i	f
0	c_i	$P(e < e^j \leftarrow c_i)$
1	c_i	1
2	c_i^0	1
2	$c_i > c_i^0$	0

line	(d_j, b)	f
1	$d_j = 0,\ b = b^0, ..., b^{j-1}, b^{\eta+1}$	1
2	$d_j = 1,\ b = b^0, ..., b^\eta$	1
3	$d_j = 2,\ b = b^{\eta+1}$	1
4	otherwise	0

line	$(d_1, ..., d_\eta, h)$	f
1	$d_i = 0,\ \forall_{j \neq i}\ d_j = 1,\ h = h^{i-1}$	1
2	$d_i = 0,\ \forall_{j \neq i}\ d_j = 1,\ h = h^i$	-1
3	$\forall_i\ d_i = 1,\ h = h^\eta$	1
4	$d_1 = 2,\ \forall_{i>1}\ d_i = 0,\ h = h^0$	-1
5	$d_1 = 2,\ \forall_{i>1}\ d_i = 0,\ h = h^\eta, h^{\eta+1}$	1
6	otherwise	0

[**MDi**] The clink and family potentials of MDi are shown in Table 2. The ilink potential of MDi is shown in Table 7.

Table 7: The ilink potential $f(d_j, b)$ of MDi and EDi

line	(d_j, b)	f
1	$d_j = 0,\ b = b^j, ..., b^{\eta+1}$	1
2	$d_j = 1,\ b = b^0, ..., b^\eta$	1
3	$d_j = 2,\ b = b^{\eta+1}$	1
4	otherwise	0

[**EDi**] The clink and ilink potentials of EDi are shown in Tables 2 and 7, respectively. The family potential of EDi is shown in Table 8.

9 Soundness and Space Complexity

Due to the existence of 4 types of gate MFs, the analysis to establish soundness of the MF of NAT models, taking into

Table 8: The family potential $f(d_1, ..., d_\eta, h)$ of EDi

line	$(d_1, ..., d_\eta, h)$	f
1	$d_i = 0,\ \forall_{j \neq i}\ d_j = 1,\ h = h^{i-1}$	-1
2	$d_i = 0,\ \forall_{j \neq i}\ d_j = 1,\ h = h^{i}$	1
3	$\forall_i\ d_i = 1,\ h = h^0$	1
4	$d_1 = 2,\ \forall_{i>1}\ d_i = 0,\ h = h^0, h^{\eta+1}$	1
5	$d_1 = 2,\ \forall_{i>1}\ d_i = 0,\ h = h^{\eta}$	-1
6	otherwise	0

account the interactions between the 4 types of gate MFs, is non-trivial. The approach we have taken is outlined below and the formal result on soundness is then stated.

To analyze the interaction between gate MFs, we decompose a gate MF into the *link layer*, consisting of clinks (including end nodes) and their potentials, and the *family layer*, consisting of the directed links and the family potential. Auxiliary variables are included in both layers.

For each gate MF, a set of desirable properties of the MPP (over its causes and auxiliary variables) from the link layer is identified, called *link potential trait*. A set of desirable properties of the MPP (over its causes and output variable) from the entire gate MF is also identified, called *output potential trait*. When a gate MF has internal input variables, its link layer is extended to include all ancestral variables, and the gate MF is extended accordingly.

For each gate MF, we show that the corresponding link potential trait holds if it has no internal input. We also show that if its internal input satisfies the output potential trait of the corresponding input gate MF, then the link potential trait also holds. We then establish that the output potential trait of the gate MF holds, no matter it has internal input or not. By integrating the per-gate MF analysis according to the rule of gate MF assignment in Table 3, the soundness of the MF of NAT models can be asserted.

Following the above approach (with more than 10 theorem-like intermediate results), we have proved the soundness as stated in Theorem 3. Due to space limitations, we omit the intermediate results and their proofs.

Theorem 3 *Let the MF be applied to a NAT model over causes $c_1, ..., c_n$ by applying MDu, EDu, MDi, and EDi to appropriate gates. Then the MPP from all potentials of the MF satisfies $f(e, c_1, ..., c_n) = P(e|c_1, ..., c_n)$.*

For the space of a NAT model, assume $m = \eta$. By Fig. 1 and Table 4, link potentials for each gate take $O(3(\eta+2)n'\eta)$ space, where n' is the number of root nodes in the gate MF graph. The family potential takes $O(3^\eta(\eta+2))$ space. If the NAT has k gates, it takes $O(3(\eta+2)n'k\ \eta + 3^\eta(\eta+2)k)$ space. The $n'k$ counts root nodes in MF graphs of all gates, and hence $n'k < 2n$. We also have $k < n$. This yields space complexity $O(n\ \eta\ (6\eta + 3^\eta))$ that is linear on n.

10 Compilation of NAT-Modeled BNs

To realize efficiency gain in BN inference by the MF of NAT models, we apply NAT modeling to BNs and compile them for LP. Consider a BN over a set V of variables with DAG D. Each root of D is assigned a prior, collected in a set PR. Each single-parent non-root is assigned a CPT, collected in a set PS. We assume that the family of each multi-parent non-root forms a NAT model, collected in a set Ψ. Then, $\Gamma = (V, D, PR, PS, \Psi)$ is a *NAT-modeled BN* (NATBN). It is compiled into a Junction Tree (JT) as follows.

Algorithm 1 *(Input: $\Gamma = (V, D, PR, PS, \Psi)$)*

```
get the skeleton $G$ of $D$ by dropping direction of links;
for each multi-parent family $\{e, c_1, ..., c_n\}$ in $D$, do
  replace subgraph of $G$ spanned by $\{e, c_1, ..., c_n\}$ with
    the MF graph $SG$ of the NAT model in $\Psi$;
  for each family in $SG$,
    connect members pairwise and drop direction of links;
triangulate $G$ into a chordal graph $G'$;
construct a JT $T$ from maximum cliques of $G'$;
assign each potential in $PR \cup PS \cup \Psi$ to a cluster in $T$;
return $T$;
```

Potentials assigned to each JT cluster are not multiplied. We refer to T as the JT of Multiplicatively Factorized NAT-modeled BN (JTMFNB), which directly supports LP (Madsen and Jensen 1999).

The above method differs from that based on MF of binary NAT models (Xiang 2012a). The gate MF for a binary NAT is a undirected tree, while the gate MF for a multi-valued NAT is hybrid and multiply connected. Hence, compilation for binary NAT-modeled BNs does not need moralization at all, while a gate level moralization is necessary in the above compilation (the inner for loop).

11 Experimental Evaluation

To evaluate soundness and space efficiency of the MF and its advantage in speeding up BN inference, a collection of 140 NATBNs are simulated, divided into 4 groups of 35 each. Each NATBN contains 100 ternary variables. For NATBNs in the same group, the number n of causes per NAT model is identically upper-bounded at 5, 7, 9 and 11, respectively. All NATBNs have the same density (5% more links than singly-connected). Each is compiled into a JTMFNB.

A *peer BN* is derived from each NATBN, where each multi-parent variable is assigned the CPT computed from the corresponding NAT model. Each peer BN is compiled into a JT representation for LP, which provides a standard for exactness and a baseline for efficiency.

For each NATBN and its peer BN, 5 randomly chosen variables are observed and posteriors for all variables are computed by LP. For each NATBN, the same posteriors are obtained from the JTMFNB and its peer BN JT, which confirms soundness of the MF empirically.

The performance is summarized in Table 9. Each row summarizes for one group of NATBNs. The space complexity of the MF analyzed in Sec. 9 does not reflect accurately the space complexity of JTMFNB, due to (1) triangulation and (2) product operation during LP. Therefore, the space efficiency of JTMFNB and peer BN JT is shown in Table 9 by the size of the state space of the JT (with sample mean and standard deviation), which is the upper bound of the actual space consumption. Time efficiency is shown by LP runtime.

Table 9: Experimental results

	Peer BN JT State Space		JTMFNB State Space		Peer BN JT Time (ms)		JTMFNB Time (ms)	
n	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$
5	11070.8	590.1	9742.7	1317.9	63.8	12.0	31.3	0.5
7	25951.4	3800.3	10546.0	1570.3	212.5	65.7	30.3	3.5
9	80061.9	6076.6	11189.8	2721.7	1117.9	749.9	33.1	7.3
11	575750.3	37149.6	10996.1	1550.2	12160.8	7658.3	30.7	2.5

As n grows from 5 to 11, peer BN JTs grow in space by 52 times, while JTMFNBs grow only 1.1 times. The runtime with peer BN JTs grows by 193 times, while inference with JTMFNBs takes about the same time. For $n = 11$, JTMFNBs use less than 2% of space as peer BN JTs, and are 396 times faster in inference. This experiment shows that the MF of NAT models allows significant improvement in space and time efficiency for sparse NAT-modeled BNs.

12 Conclusion

The main contribution of this work is the formulation of multiplicative factorization for multi-valued NAT models. In comparison with the MF for binary NAT models, the proposed MF is a hybrid and multiply connected graphical model. To enable sound and space efficient factorization, we have shown that multiple alternative gate MFs are necessary. As the result, parameterization of these gate MFs, their integration, and the overall analysis for soundness become significantly more sophisticated than the binary case. Overcoming these challenges, we have shown that the formulated MF is exact and is space efficient (linear complexity on the number of causes per effect).

In addition, we have shown that a multi-valued, dual NIN-AND gate is equivalent to noisy-MAX. Based on the proposed MF for multi-valued NAT models, we extended the scheme for LP with binary NATBNs to multi-valued NATBNs. This scheme is more powerful than LP based on MF of noisy-MAX (Madsen and D'Ambrosio 2000), since NATBNs are strictly more expressive than noisy-MAX modeled BNs as implied by the above result about noisy-MAX. We experimentally demonstrated that JTMFNBs compiled from sparse NATBNs allow exact but significantly more efficient LP both in space and in time. Although binarization (Antonucci et al. 2006) allows MF of binary NAT models to apply directly to multi-valued BNs, it can be shown to be less space efficient than the proposed MF. Behavior of the proposed MF when NAT models degrade into binary will be examined in future work.

In summary, this work opens a promising direction along which significantly less computational resource is necessary for probabilistic reasoning with general BNs, making them deployable in pervasive computing devices.

Acknowledgement

Financial support from NSERC, Canada is acknowledged.

References

Antonucci, A.; Zaffalon, M.; Ide, J.; and Cozman, F. 2006. Binarization algorithms for approximate updating in credal nets. In *Frontiers in Artificial Intelligence and Applications*, volume 142, 120–131. IOS Press.

Diez, F. 1993. Parameter adjustment in Bayes networks: The generalized noisy OR-gate. In Heckerman, D., and Mamdani, A., eds., *Proc. 9th Conf. on Uncertainty in Artificial Intelligence*, 99–105. Morgan Kaufmann.

Henrion, M. 1989. Some practical issues in constructing belief networks. In Kanal, L.; Levitt, T.; and Lemmer, J., eds., *Uncertainty in Artificial Intelligence 3*. Elsevier Science Publishers. 161–173.

Jensen, F.; Lauritzen, S.; and Olesen, K. 1990. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly* (4):269–282.

Madsen, A., and D'Ambrosio, B. 2000. A factorized representation of independence of causal influence and lazy propagation. *Inter. J. Uncertainty, Fuzziness and Knowledge-Based Systems* 8(2):151–166.

Madsen, A., and Jensen, F. 1999. Lazy propagation: A junction tree inference algorithm based on lazy evaluation. *Artificial Intelligence* 113(1-2):203–245.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Savicky, P., and Vomlel, J. 2007. Exploiting tensor rank-one decomposition in probabilistic inference. *Kybernetika* 43(5):747–764.

Takikawa, M., and D'Ambrosio, B. 1999. Multiplicative factorization of noisy-max. In *Proc. 15th Conf. Uncertainty in Artificial Intelligence*, 622–630.

Xiang, Y., and Jia, N. 2007. Modeling causal reinforcement and undermining for efficient CPT elicitation. *IEEE Trans. Knowledge and Data Engineering* 19(12):1708–1718.

Xiang, Y. 2012a. Bayesian network inference with NIN-AND tree models. In Cano, A.; Gomez-Olmedo, M.; and Nielsen, T., eds., *Proc. 6th European Workshop on Probabilistic Graphical Models*, 363–370.

Xiang, Y. 2012b. Non-impeding noisy-and tree causal models over multi-valued variables. *International J. Approximate Reasoning* 53(7):988–1002.

Zhang, N., and Poole, D. 1996. Exploiting causal independence in bayesian network inference. *J. Artificial Intelligence Research* 5:301–328.