Trans-Causalizing NAT-Modeled Bayesian Networks

Yang Xiang and Dylan Loker, School of Computer Science, University of Guelph, Canada

***Abstract*—Conditional independence encoded in Bayesian networks avoids combinatorial explosion on the number of variables. However, Bayesian networks are still subject to exponential growth of space and inference time on the number of causes per effect variable in conditional probability tables. A number of space-efficient local models exist that allow efficient encoding of dependency between an effect and its causes, and can also be exploited for improved inference efficiency. We focus on the Non-Impeding Noisy-AND Tree (NIN-AND Tree or NAT) models because of multiple merits. We present a novel framework, *trans-causalization* of NAT-modeled Bayesian networks, by which causal independence embedded in NAT models is exploited for more efficient inference. We show that trans-causalization is exact and yields polynomial space complexity. We demonstrate significant efficiency gain on inference based on lazy propagation and sum-product networks. (Keywords: Bayesian networks, Causal independence models, Probabilistic inference)**

I. Introduction

Bayesian networks [1] have been widely used in intelligent systems for inference in complex, partially observable, and stochastic environments, e.g., [2]. A discrete Bayesian network (BN) encodes probabilistic knowledge about an application environment whose states are described by a set V of variables. Traditional probabilistic approach encodes such knowledge by a joint probability distribution (JPD) over V. Since the JPD has a size exponential in $|V|$, inference with a JPD is intractable when $|V|$ is large. BNs overcome this difficulty by encoding the dependency and conditional independency among variables through a directed acyclic graph (DAG). Relative to the DAG, we shall refer to a child variable and its parents as an effect and its causes[1]. The dependency of an effect on its causes in the DAG is quantified by a conditional probability table (CPT). The CPT specifies how probable the effect takes on a value, when the causes take on each possible configuration. Once the DAG and the CPTs (one per variable) are specified, the JPD over V is uniquely defined. The total number of CPTs in a BN is $|V|$, and each CPT has a size exponential in the size of the variable family (the effect plus its causes). Hence, when each family is small, the number of probability parameters in a BN is significantly less than a JPD. By exploiting conditional independence encoded in the DAG, inference with BNs can be performed much more efficiently than using JPDs.

However, when a BN contains large families, the exponential parameter growth occurs at the family level, and can diminish the computational advantage of BNs over JPDs. To address this issue, a number of space-efficient local models have been proposed. Rather than quantifying the dependency of an effect on its causes by a CPT, how causes interact in causing the effect is characterized, which allows quantification of the dependency by specifying probability parameters that are much less than exponential in the family size. They include noisy-OR [1], noisy-MAX [3], [4], context-specific independence (CSI) [5], noisy-AND [6], recursive noisy-OR [7], Non-Impeding Noisy-AND Tree (NIN-AND Tree or NAT) [8], [9], DeMorgan [10], tensor-decomposition [11], cancellation model [12], among others. Among these local models, noisy-OR, noisy-MAX, noisy-AND, recursive noisy-OR, DeMorgan, and NAT belong to a class of causal independency models (CIMs). They assume that the strength of a cause in causing the effect to occur is independent of whether other causes of the effect are active.

This work focuses on representing BN CPTs through NAT models [8] due to several merits: First, causes of the same effect may interact qualitatively differently by reinforcing or undermining each other. Widely utilized models such as noisy-OR, noisy-AND, and noisy-MAX can express only reinforcing interactions, while NAT models can express both reinforcing and undermining causal interactions.

Second, reinforcing and undermining may be between individual causes, or between subsets of causes. DeMorgan model can express both reinforcing and undermining, but only between individual causes, while NAT models can express these interactions both between individual causes and between subsets of causes. For instance, a subset of causes may reinforce each other and so do the causes in another subset, but the two subsets may undermine each other. NAT models can express such interaction while DeMorgan cannot. Furthermore, NAT models can express such mixture of reinforcing and undermining recursively among subsets of causes.

Third, noisy-OR and noisy-AND are applicable to binary variables only, while NAT models are applicable to multi-valued variables. Fourth, most other CIMs have been defined over ordinal variables (also referred to as graded variables). NAT models are defined over both ordinal and nominal variables [13]).

Fifth, NAT models strictly generalize noisy-OR, noisy-MAX [14], and DeMorgan [8], and at the same time have the same space complexity. That is, the number of parameters required by a NAT model is linear in the number of causes, which is the same as noisy-OR, noisy-MAX, and DeMorgan. In Section X, we discuss several representational issues that relate NAT models with other CIMs.

Sixth, reinforcing or undermining causal interactions that NAT models express are inequality based, while CSI are equality based. Hence, NAT models are orthogonal and hence complimentary to CSI.

Seventh, through multiplicative factorization (MF) of NAT-modeled BNs, inference based on lazy propagation (LP) [15] can be two orders of magnitude faster for a range of sparse BNs [14], [16]. A NAT-modeled BN is a BN where each multi-

[1]We use causality loosely and interchangeably with asymmetric dependency.

parent family is represented as a NAT model. The sparseness is signified by the lower number of arcs beyond a tree DAG. Inference complexity of a BN is exponential in its tree-width. The tree-width of a BN is one less than the size of the largest cluster, when its DAG structure is transformed into a best junction tree. Inference with a NAT-modeled BN is significantly more efficient than the equivalent BN based on tabular CPTs, when its DAG structure is sparse but has a large tree-width (due to large variable families).

Although MF has achieved impressive performance in improving the efficiency of inference in NAT-modeled BNs, it has a limitation. A NAT model (encoding a BN CPT) consists of several NIN-AND gates. The MF of each gate has one numerical potential (among others) that is exponential in the domain size of the effect variable. In this work, we develop a framework alternative to MF, referred to as *trans-causalization* of NAT-modeled BNs, which eliminates such exponential components.

In general, trans-causalized BNs are homogenous in representation: All variables have small families (no more than 2 parents) and hence a small tabular CPT. This eliminates heterogeneity of NAT-modeled BNs, where CPTs of large families are encoded by NAT models. It allows inference with trans-causalized BNs to be conducted using any BN inference algorithm, while significantly reducing the space complexity. As the result, the space complexity of trans-causalized BNs is polynomial. At the same time, trans-causalized BNs preserve exactly the probability distributions of NAT-modeled BNs. We demonstrate superior inference performance of trans-causalized BNs through inference by LP and inference by sum-product networks (SPNs) [17], [18], [19].

A significant part of the effort (and contribution) of this work is to rigorously establish the exactness of trans-causalized BNs. This effort is complicated by the rich structural details of NAT models (alternative gate types, their alternations in the NAT tree, and their levels in the NAT tree, etc.) that are necessary for their expressiveness. To that end, the paper applied formal analysis and proofs. Due to space limit, proofs of most formal results (theorems and lemmas) are included in Supplementary Materials.

The remainder is organized as follows. Section II reviews the background on NAT modeling. This is followed by the motivation of this research and an overview of main computational steps of trans-causalization in Section III. In Sections IV and V, we present how to trans-causalize dual and direct NIN-AND gate models. Further reduction of tree-width of the trans-causalized representation is presented in Section VI. The trans-causalization is extended to general NAT models in Section VII and to NAT-modeled BNs in Section VIII. The impact of trans-causalization is empirically evaluated in Section IX.

II. Background on NAT Models

This section briefly reviews background on NAT models. More details can be found in [8], [13]. A NAT model is defined over an effect e and a set of n causes $C = \{c_1, ..., c_n\}$ that are multi-valued, where $e \in D_e = \{e^0, ..., e^\eta\}$ $(\eta \geq 1)$ and $c_i \in \{c_i^0, ..., c_i^{m_i}\}$ $(i = 1, ..., n, m_i \geq 1)$. C and e form one family (a child variable plus its parents) in a BN, whose dependence is quantified by a tabular CPT by default. Values e^0 and c_i^0 are *inactive*. Other values (may be written as e^+ or c_i^+) are *active*. A higher index often means higher intensity (graded or ordinal variables), but that is not necessary (see [13] for generalization to nominal variables).

A causal event is a *success* or *failure* depending on if e is active up to a given value, is *single-* or *multi-causal* depending on the number of active causes, and is *simple* or *congregate* depending on value range of e. 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 probability of a *simple single-causal success*, and

$$P(e \geq e^k \leftarrow c_1^{j_1}, ..., c_q^{j_q}) = P(e \geq e^k | c_1^{j_1}, ..., c_q^{j_q}, c_z^0 : c_z \in C \setminus X)$$

is probability of a *congregate multi-causal success*, where $j_1, ..., j_q > 0$, $X = \{c_1, ..., c_q\}$ $(q > 1)$. The latter may be denoted as $P(e \geq e^k \leftarrow \underline{x}^+)$. Interactions among causes may be reinforcing or undermining as defined below.

Definition 1: Let e^k be an active effect value, $R = \{W_1, ..., W_\omega\}$ $(\omega \geq 2)$ be a partition of a set $X \subseteq C$ of causes, $S \subset R$, and $Y = \cup_{W_i \in S} W_i$. Sets of causes in R reinforce each other relative to e^k, iff $\forall S \; P(e \geq e^k \leftarrow \underline{y}^+) \leq P(e \geq e^k \leftarrow \underline{x}^+)$. They undermine each other iff $\forall S \; P(e \geq e^k \leftarrow \underline{y}^+) > P(e \geq e^k \leftarrow \underline{x}^+)$.

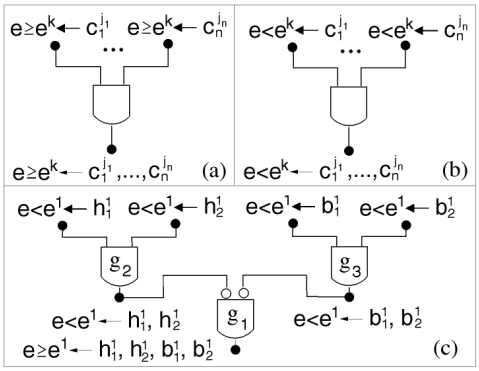

Fig. 1. (a) Direct NIN-AND gate. (b) Dual NIN-AND gate. (c) NAT.

A NAT consists of multiple NIN-AND gates. A *direct* gate involves disjoint sets of causes $W_1, ..., W_\omega$. Each input event is a success $e \geq e^k \leftarrow \underline{w}_i^+$ $(i = 1, ..., \omega)$, e.g., Fig. 1 (a) where each W_i is a singleton. The output event is $e \geq e^k \leftarrow \underline{w}_1^+, ..., \underline{w}_\omega^+$. The probability of output event of a direct NIN-AND gate is

$$P(e \geq e^k \leftarrow \underline{w}_1^+, ..., \underline{w}_\omega^+) = \prod_{i=1}^{\omega} P(e \geq e^k \leftarrow \underline{w}_i^+). \quad (1)$$

Direct gates encode undermining causal interactions. Each input event of a *dual* gate is a failure $e < e^k \leftarrow \underline{w}_i^+$, e.g., Fig. 1 (b). The output event is $e < e^k \leftarrow \underline{w}_1^+, ..., \underline{w}_\omega^+$. The probability of output event of a dual NIN-AND gate is

$$P(e < e^k \leftarrow \underline{w}_1^+, ..., \underline{w}_\omega^+) = \prod_{i=1}^{\omega} P(e < e^k \leftarrow \underline{w}_i^+). \quad (2)$$

Dual gates encode reinforcement causal interactions.

A NAT encodes complex causal interactions beyond a single direct or dual gate. For example, consider a surface enhancer application. Acidic enhancers h_1 and h_2 are more effective when both are applied. Basic enhancers b_1 and b_2 work similarly. However, when enhancers from both groups are combined, their effectiveness is reduced. Formally, h_1, h_2, b_1, and b_2 are causes, and surface enhancement is their effect e. Causes h_1 and h_2 reinforce each other, and so do b_1 and b_2. However, the two groups undermine each other. The NAT in Fig. 1 (c) encodes their causal interactions.

To quantify causal strength of each cause, probabilities of simple single-causal success, $P(e^k \leftarrow c_i^j)$ $(j,k>0)$, also called *single-causals*, are specified. Suppose that surface enhancer h_1 has 2 alternative grades $h_1 = h_1^1$ and $h_1 = h_1^2$. We have $h_1 \in \{h_1^0, h_1^1, h_1^2\}$, where h_1^0 expresses that h_1 enhancer is not applied. Similarly, suppose that we have $h_2 \in \{h_2^0, h_2^1, h_2^2\}$, $b_1 \in \{b_1^0, b_1^1, b_1^2\}$, $b_2 \in \{b_2^0, b_2^1, b_2^2\}$, and $e \in \{e^0, e^1, e^2\}$. We need to specify $2 \times 2 \times 4 = 16$ single-causals. From Fig. 1 (c) and single-causals, $P(e \geq e^1 \leftarrow h_1^1, h_2^1, b_1^1, b_2^1)$ can be obtained. From the single-causals and all derivable NATs [20], CPT $P(e|h_1, h_2, b_1, b_2)$ is uniquely specified [8]. A NAT model is specified by the topology and a set of single-causals with a space linear in n.

A BN, where the CPT of every family of size 3 or larger is a NAT model, is a *NAT-modeled BN*. A discrete BN where every CPT is tabular has a space complexity of $O(N\ \kappa^n)$, where N is the number of variables, κ is the size of largest variable domains, and n is the largest number of parents per variable. On the other hand, a NAT-modeled BN has a linear space complexity of $O(N\ \kappa\ n)$. The space efficiency of NAT-modeled BNs can extend to efficiency in inference time through MF [14]. With MF of NAT-modeled BNs, inference based on LP can be two orders of magnitude faster for a range of sparse BNs [14], [16].

Before closing this section, we demonstrate, using the enhancer example, that while NAT models have the same linear space complexity as other CIMs, they are more expressive. In particular, we compare with noisy-OR, noisy-AND, noisy-MAX, and DeMorgan, relative to the 1st five advantages of NAT models stated in Section I, which we refer below as the 1st point, the 2nd point, etc. To do so, we first specify the 16 single-causals. Although they generally differ, for simplicity, we assume following values for cause $c = h_1, h_2, b_1, b_2$:

$$P(e^1 \leftarrow c^1) = 0.3, \quad P(e^2 \leftarrow c^1) = 0.4,$$
$$P(e^1 \leftarrow c^2) = 0.4, \quad P(e^2 \leftarrow c^2) = 0.5.$$

These single-causals and the NAT in Fig. 1 (c) allow computation of a unique NAT model CPT $P(e|h_1, h_2, b_1, b_2)$.

On the 1st point, NAT model CPT has

$$P(e \geq e^2|h_1^1, h_2^0, b_1^0, b_2^0) = 0.4, \quad P(e \geq e^2|h_1^0, h_2^1, b_1^0, b_2^0) = 0.4,$$
$$P(e \geq e^2|h_1^1, h_2^1, b_1^0, b_2^0) = 0.64,$$

expressing reinforcing between h_1 and h_2. NAT CPT also has

$$P(e \geq e^2|h_1^1, h_2^0, b_1^0, b_2^0) = 0.4, \quad P(e \geq e^2|h_1^0, h_2^0, b_1^1, b_2^0) = 0.4,$$
$$P(e \geq e^2|h_1^1, h_2^0, b_1^1, b_2^0) = 0.16,$$

expressing undermining between h_1 and b_1. On the other hand, noisy-OR, noisy-AND, and noisy-MAX can only express reinforcing, but not undermining.

On the 2nd point, NAT model CPT expresses reinforcing between h_1 and h_2 as above, and has

$$P(e \geq e^2|h_1^0, h_2^0, b_1^1, b_2^0) = 0.4, \quad P(e \geq e^2|h_1^0, h_2^0, b_1^0, b_2^1) = 0.4,$$
$$P(e \geq e^2|h_1^0, h_2^0, b_1^1, b_2^1) = 0.64,$$

expressing reinforcing between b_1 and b_2. NAT CPT also has

$$P(e \geq e^2|h_1^1, h_2^1, b_1^1, b_2^1) = 0.4096,$$

expressing undermining between both groups. On the other hand, DeMorgan cannot express such complex mixture of causal interactions.

On the 3rd point, since all enhancer causes and the effect are ternary, noisy-OR and noisy-AND cannot model this application. On the 4th point, the enhancer NAT model does not require that variables be graded, e.g., $h_1^0 \prec h_1^1 \prec h_1^2$. To do the some, alternative CIMs need to adopt the paradigm of NAT models in [13]. On the 5th point, if we limit enhancers to h_1 and h_2 only, then the restricted application can be modeled by noisy-MAX, and DeMorgan. If we further restrict h_1, h_2, and e to binary, then it can be modeled by noisy-OR. If we limit enhancers to h_1 and b_1 only, then it can be modeled by DeMorgan. However, the enhancer application, as we specified, cannot be modeled by any of noisy-OR, noisy-MAX, and DeMorgan.

III. Trans-Causalization: Motivation & Overview

Although MF has achieved impressive inference performance in NAT-modeled BNs, it has a limitation. The MF of a NAT model over an effect e converts each NIN-AND gate into a hybrid network segment with both directed and undirected links. Nodes in the segment are located at 3 levels: 0, 1, and 2. Level 0 consists of a single variable x (corresponding to e). Level 1 consists of variables one per active value of e (a total of $|D_e| - 1$). Each of them is connected to x by a directed link. Hence, the set π of nodes at level 1 are parents of x and $\{x\} \cup \pi$ form a family. Level 2 consists of variables one per input event of the NIN-AND gate. Each node at level 2 is connected to each node in level 1 by a undirected link. There are no links between nodes at the same level, and nor between level 2 and 0. Each undirected link between levels 2 and 1 is quantified by a link potential. The single family is quantified by a potential over $\{x\} \cup \pi$. Since $|\pi|$ is linear in $|D_e|$, the family potential is exponential in $|D_e|$.

We present a framework alternative to MF, which does not suffer from such exponential components. In NAT models, input and output of each gate are causal events (Fig. 1 (c)). The new framework *transforms* casual events in the NAT model into regular variables (as those in a BN). Hence, we refer to the framework as *trans-causalization*.

The framework consists of three levels of computations: gate level, NAT level, and BN level. The gate level computation converts each NIN-AND gate into a BN segment by introducing a probabilistic auxiliary variable for each input event of the gate. These auxiliary variables become children of the

causes in the input events, and parents of the effect variable. The CPT of each auxiliary variable is probabilistic, while the CPT of the effect is deterministic.

The NAT level computation operates on each NAT Model over a BN family. After a BN segment is created for each NIN-AND gate in the NAT, the segments are merged into the BN segment of the NAT model, such that it encodes CPT of the NAT model exactly. If an NIN-AND gate feeds into another in the NAT, its effect variable is replaced with a *quasi-effect* variable. To ensure exactness of the BN segment CPT while maximizing efficiency, the domain of quasi-effect may differ from that of effect e, depending on location of the gate in the NAT. CPT of the quasi-effect is altered accordingly.

The BN level computation operates on a NAT-modeled BN. For each BN family over an effect e and parents $c_1,...,c_n$ whose dependency is encoded as a NAT model, create the BN segment as above. Delete the link from each c_i to e in the BN, and reconnect the family by the BN segment of the NAT model. The resultant is a trans-causalized BN, which encodes the JPD of the NAT-modeled BN exactly.

In subsequent sections, we elaborate the computation at each level, and establishes the exactness.

IV. Trans-Causalization of Dual Gate Models

First, we trans-causalize a dual NIN-AND gate model, a building block of NAT models, into a BN segment whose general BN structure is shown in Fig. 2 (a). The leaf node is the effect variable e. The root nodes are cause variables divided into ω groups, according to the cause partition $W_1,...,W_\omega$ of the dual gate model. Each group W_i forms the parent set of a probabilistic auxiliary variable z_i with domain D_e. The z_i represents impact of W_i to effect e.

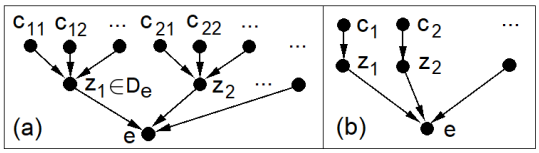

Fig. 2. (a) General DAG structure of a SBDu (see Def. 2), where the parent set of z_1 is W_1. (b) When each z_i has a single parent ($\theta = 1$).

Each W_i includes $\theta_i \geq 1$ causes: $W_i = \{c_{i1},...,c_{i\theta_i}\}$. We denote $\theta = \max_i \theta_i$. When $\theta = 1$, each W_i is a singleton, the structure degenerates to Fig. 2 (b), and we refer to the BN segment as the segment with $\theta = 1$. The CPT at z_i is

$$P(z_i|c_{i1},...,c_{i\theta_i}) = \begin{cases} 1, & if\ z_i = e^0\ and\ \forall_x\ c_{ix} = c_{ix}^0, \\ P(e^j \leftarrow c_{i1},...,c_{ik}), & if\ z_i = e^j,\ j > 0, k \geq 1, \\ & c_{i1},...,c_{ik}\ are\ active,\ and \\ & c_{i,k+1},...,c_{i\theta_i}\ are\ inactive. \end{cases} \quad (3)$$

The 1st formula says that when all causes in the ith group are inactive, they cannot render e active. The 2nd formula expresses the impact to e when some causes in the ith group are active, where $P(e^j \leftarrow c_{i1},...,c_{ik})$ is from the dual gate model. Note that given $P(e^j \leftarrow c_{i1},...,c_{ik})$ for $j > 0$, the value for $j = 0$ is uniquely derived. When $\theta = 1$, the CPT at z_i becomes the *single-causal* (SC) CPT:

$$P(z_i|c_i) = \begin{cases} 1, & if\ z_i = e^0\ and\ c_i = c_i^0, \\ P(e^j \leftarrow c_i), & if\ z_i = e^j, j > 0,\ and\ c_i > c_i^0. \end{cases} \quad (4)$$

A gate in a NAT has $\theta = 1$, if there exists no upstream gate in the NAT. However, whenever a gate has one or more upstream gates, some input event of the gate will involve a subset W_i of upstream causes, and therefore $\theta > 1$. Hence, the generality of $\theta \geq 1$ is needed in the analysis of trans-causalization of NAT models. This is the case for all BN segments presented below.

The CPT at e, referred to as a *MAX* CPT, encodes a *MAX* function, where the domain of every variable is D_e and the number of α_i variables is finite:

$$P(\tau|\alpha_1,\alpha_2,...) = \begin{cases} 1, & if\ \tau = MAX(\alpha_1,\alpha_2,...), \\ 0, & otherwise. \end{cases} \quad (5)$$

For the MAX CPT at e, τ is substituted by e and $\alpha_1,\alpha_2,...$ by $z_1,...,z_\omega$.

Definition 2: Given a dual NIN-AND gate model over e and $C = W_1 \cup ... \cup W_\omega$, let G be the DAG derived as Fig. 2, and CP be the set of CPTs specified by Eqns. (3) and (5). Then $\Phi = (C, e, G, CP)$ is the *BN Segment Base for the Dual gate model*, or ***SBDu***.

Theorem 1 shows that the SBDu is probabilistically equivalent to the deriving dual gate model, e.g., that in Fig. 1 (b).

Theorem 1: Let $\Phi = (C, e, G, CP)$ be the SBDu of a dual NIN-AND gate model. Then the CPT $P_\Phi(e|W_1,...,W_\omega)$ defined by marginalized product

$$\sum_{z_1,...,z_\omega} \left(P_\Phi(e|z_1,...,z_\omega) \prod_{i=1}^{\omega} P_\Phi(z_i|c_{i1},...,c_{i\theta_i}) \right)$$

is identical to that of the dual gate model.

As presented, Theorem 1 is applicable to the SBDu where $\theta \geq 1$. When $\theta = 1$, it says that the SBDu is equivalent to a dual gate model such as that in Fig. 1 (b). Note that events in Fig. 1 are causal events, while variables in Fig. 2 are regular variables. Hence, the name *trans-causalization*.

V. Trans-Causalization of Direct Gate Models

Next, we trans-causalize a direct NIN-AND gate model, the other building block of NAT models, into a BN segment. The structure of the BN segment is similar to Fig. 2, with cases $\theta > 1$ and $\theta = 1$ illustrated in Fig. 3 (a) and (b), respectively. However, the domain of each z_i is $D_a = \{e^0,...,e^\eta, aaci\}$,

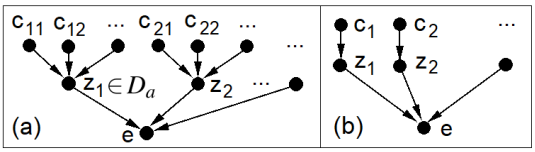

Fig. 3. (a) General DAG structure of a SBDi. (b) When $\theta = 1$.

where an extra value *aaci* (all above causes inactive) is added to D_e. The value *aaci* is used to signify that all cause

parents of z_i are inactive. It is necessary for implementing the non-impeding behavior of direct gate (inactive causes do not impede the function of active causes), as is evident in Eqn. (8) below. When values of z_i are compared, the notational convention $e^0 < ... < e^\eta < aaci$ is assumed.

The CPT at z_i is

$$P(z_i|c_{i1},...,c_{i\theta_i}) = \begin{cases} 1, & if\ z_i = aaci\ and\ \forall_x\ c_{ix} = c_{ix}^0, \\ P(e^j \leftarrow c_{i1},...,c_{ik}), & if\ z_i = e^j, j > 0, k \geq 1, \\ & c_{i1},...,c_{ik}\ are\ active,\ and \\ & c_{i,k+1},...,c_{i\theta_i}\ are\ inactive. \end{cases} \quad (6)$$

The 1st formula explicitly signifies that all causes in W_i are inactive. The 2nd formula covers cases where some causes are active. When $\theta = 1$, the CPT at z_i becomes a *single-causal-plus* (SC^+) CPT, where $^+$ signifies the enlarged domain of z_i beyond D_e:

$$P(z_i|c_i) = \begin{cases} 1, & if\ z_i = aaci\ and\ c_i = c_i^0, \\ P(e^j \leftarrow c_i), & if\ z_i = e^j, j > 0,\ and\ c_i > c_i^0. \end{cases} \quad (7)$$

The CPT at e, referred to as *PMIN* CPT, encodes a pseudo-MIN (PMIN) function over a finite set of arguments, where each argument has domain D_a (hence, pseudo) and the function range is D_e:

$$PMIN(\alpha_1, \alpha_2, ...) = \begin{cases} e^0, & if\ \forall_i \alpha_i = aaci, \\ MIN(\alpha'_1,...,\alpha'_m), & if\ \alpha'_1,...,\alpha'_m \neq aaci\ (m > 0). \end{cases}$$

The PMIN CPT at e is the following:

$$P(\tau|\alpha_1, \alpha_2, ...) = \begin{cases} 1, if\ \forall_i\ \alpha_i = aaci \wedge \tau = e^0, \\ 1, if \alpha'_1,...,\alpha'_m \neq aaci\ (m > 0) \wedge \tau = MIN(\alpha'_1,...,\alpha'_m). \end{cases} \quad (8)$$

For the PMIN CPT at e, τ is substituted by e and $\alpha_1, \alpha_2, ...$ by $z_1, ..., z_\omega$. We define the BN segment below and establish its soundness.

Definition 3: Given a direct NIN-AND gate model over e and $C = W_1 \cup ... \cup W_\omega$, let G be the DAG derived as Fig. 3, and CP be the set of CPTs specified by Eqns. (6) and (8). Then $\Phi = (C, e, G, CP)$ is the *BN Segment Base for the Direct gate model*, or ***SBDi***.

Theorem 2 shows that SBDi is probabilistically equivalent to the deriving direct gate model, e.g., that in Fig. 1 (a).

Theorem 2: Let $\Phi = (C, e, G, CP)$ be the SBDi of a direct NIN-AND gate model. Then the CPT $P_\Phi(e|W_1,...,W_\omega)$ defined by marginalized product

$$\sum_{z_1,...,z_\omega} \left(P_\Phi(e|z_1,...,z_\omega) \prod_{i=1}^{\omega} P_\Phi(z_i|c_{i1},...,c_{i\theta_i}) \right)$$

is identical to that of the direct gate model.

Towards the end of proof, if the domain of z_i does not contain value *aaci*, we would have $\sum_{z_i \geq e^k} P_\Phi(z_i|\underline{w}_i^0) = 0$ for $k > 0$ and $i = m+1, ..., \omega$. As the result, the related product would be zero (impeding) and be incorrect. This demonstrates that the standard noisy-MIN is insufficient. Instead, to realize the non-impeding behavior of the direct NIN-AND gate model, the enlarged domain of z_i and the PMIN CPT are necessary.

Theorem 2 is applicable to the SBDi where $\theta \geq 1$. When $\theta = 1$, it says that the SBDi is equivalent to a direct gate model such as that in Fig. 1 (a).

VI. Reduction of Tree-Width for BN Segments

The cost of inference using a BN is critically dependent on its tree-width. By reducing tree-widths of BN segments, the overall tree-width of a BN may also be reduced. We consider the SBDu first, followed by the SBDi.

A SBDu with $\theta = 1$ has a tree-width of ω (since the tree-width is one less than the largest cluster size in the best junction tree, and the cluster size is $\omega + 1$). Below, we take advantage of the deterministic CPT $P(e|z_1,...,z_\omega)$ by Eqn. (4), and apply divorcing [21] to reduce tree-width of the segment from ω to 2. Fig. 4 shows the modified DAG structure. A total of $\omega - 2$ deterministic auxiliary variables y_i are introduced. When $\theta = 1$ (Fig. 4 (b)), each node has no more than two parents, and its tree-width is 2.

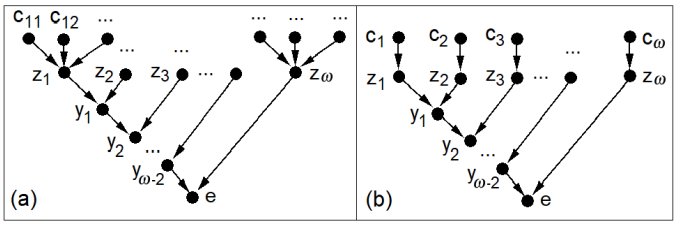

Fig. 4. (a) General DAG structure of a divorcing segment. (b) When $\theta = 1$.

DSDu We refer to the modified SBDu as the *Divorcing BN Segment for the Dual gate model*, or ***DSDu***. The domain of each z_i and each y_j is D_e. The CPT at each z_i is specified by Eqn. (3). The CPT at e and each y_i ($i = 1, ..., \omega - 2$) is a MAX CPT defined by Eqn. (5).

Assume that all cause variables have domain size $\eta + 1$ as e. The CPT $P(e|z_1,...,z_\omega)$ of a SBDu has a size of $(\eta+1)^{\omega+1}$. On the other hand, for the DSDu, the total size of all CPTs at e and each y_i ($i = 1, ..., \omega - 2$) is $(\omega - 1)(\eta + 1)^3$. For $\omega = \eta = 4$, the two sizes are 3125 and 375, respectively.

Although divorcing yields significant space efficiency, it remains to be shown that $P(e|z_1,...,z_\omega)$ defined by the CPTs at e and each y_i ($i = 1, ..., \omega - 2$) is equivalent to that of the SBDu. Therefore, we formally justify the equivalence below, through several lemmas and a theorem. The formal analysis is extended later to analyze the modified SBDi.

In Lemmas 1 through 4, Φ denotes a SBDu, with MAX CPT $P_\Phi(e|z_1,...,z_\omega)$. Ψ denotes the DSDu, associated with CPT

$$P_\Psi(e|z_1,...,z_\omega) = \sum_{y_1,...,y_{\omega-2}} Q(z_1,...,z_\omega, y_1,...,y_{\omega-2}, e), \quad (9)$$

where the function $Q(.)$ is

$$Q(z_1,...,z_\omega, y_1,...,y_{\omega-2}, e) = P_\Psi(y_1|z_1,z_2) \left(\prod_{i=2}^{\omega-2} P_\Psi(y_i|y_{i-1}, z_{i+1}) \right) P_\Psi(e|y_{\omega-2}, z_\omega).$$

Lemmas 1 and 2 establish conditions where $Q(.) = 1$ and $Q(.) = 0$. Lemmas 3 and 4 justify when probability $P_\Psi(e|z_1,...,z_\omega) = 1$ and when $P_\Psi(e|z_1,...,z_\omega) = 0$.

Lemma 1: For any tuple $t = (z_1,...,z_\omega,y_1,...,y_{\omega-2},e)$, $Q(t) = 1$ iff $y_1 = MAX(z_1,z_2)$, $\cdots, y_{\omega-2} = MAX(z_1,...,z_{\omega-1})$, and $e = MAX(z_1,...,z_\omega)$.

Lemma 2 shows the condition where $Q(.) = 0$.

Lemma 2: For any tuple $t = (z_1,...,z_\omega,y_1,...,y_{\omega-2},e)$, we have $Q(t) = 0$ if one of the following does not hold:

$$y_1 = MAX(z_1,z_2),\ ...,\ y_{\omega-2} = MAX(z_1,...,z_{\omega-1}),$$

$$e = MAX(z_1,...,z_\omega).$$

The condition where probability $P_\Psi(e|z_1,...,z_\omega) = 1$ is established in Lemma 3.

Lemma 3: For any tuple $(z_1,...,z_\omega,e)$, we have $P_\Psi(e|z_1,...,z_\omega) = 1$, if and only if $e = MAX(z_1,...,z_\omega)$.

Lemma 4 justifies when probability $P_\Psi(e|z_1,...,z_\omega) = 0$.

Lemma 4: For any tuple $(z_1,...,z_\omega,e)$ such that $e \neq MAX(z_1,...,z_\omega)$, we have $P_\Psi(e|z_1,...,z_\omega) = 0$.

Theorem 3 establishes that a DSDu is probabilistically equivalent to the deriving dual gate model.

Theorem 3: Let Ψ be the DSDu of a dual NIN-AND gate model, and $P_\Psi(e|z_1,...,z_\omega)$ be defined by CPTs in Ψ. Then the CPT $P_\Psi(e|W_1,...,W_\omega)$ defined by marginalized product

$$\sum_{z_1,...,z_\omega} \left(P_\Psi(e|z_1,...,z_\omega) \prod_{i=1}^{\omega} P_\Psi(z_i|c_{i1},...,c_{i\theta_i}) \right)$$

equals $P(e|W_1,...,W_\omega)$ of the dual gate model.

DSDi Next, we apply divorcing to the SBDi. We refer to the modified segment as the *Divorcing BN Segment for the Direct gate model*, or ***DSDi***. Its DAG structure is the same as Fig. 4. However, the domain of each z_i and each y_j is D_a (including *aaci*). The CPT $P(e|z_1,...,z_\omega)$ of a SBDi has a size of $(\eta+1)(\eta+2)^\omega$. For the DSDi, the total size of all CPTs at e and each y_i $(i = 1,...,\omega-2)$ is $(\omega-2)(\eta+2)^3 + (\eta+1)(\eta+2)^2$. For $\omega = \eta = 4$, the two sizes are 6480 and 612, respectively.

The CPT at each z_i is the same as SBDi, i.e., by Eqn. (6). The CPT at e is a PMIN CPT defined by Eqn. (8), where condition variables are $y_{\omega-2}$ and z_ω. When one of $y_{\omega-2}$ and z_ω is not *aaci*, the MIN function is trivialized.

The CPT at each y_i $(i = 1,...,\omega-2)$, referred to as a $PMIN^+$ CPT, encodes the pseudo-MIN-plus (PMIN^+) function:

$$PMIN^+(\alpha_1,\alpha_2) = \begin{cases} aaci, & if\ \alpha_i = aaci\ (i = 1,2), \\ MIN(\alpha_1',\alpha_m'), & if\ \alpha_1',\alpha_m' \neq aaci\ (m > 0), \end{cases}$$

where $+$ signifies function range with *aaci*. When $m = 1$, the MIN function is trivial. The PMIN^+ CPT at each y_i is the following, where τ is substituted by y_i, and α_i are substituted by parents of y_i:

$$P(\tau|\alpha_1,\alpha_2) = \begin{cases} 1, & if\ \alpha_i = aaci\ (i = 1,2) \wedge \tau = aaci, \\ 1, & if\ \alpha_1',\alpha_m' \neq aaci\ (m > 0) \wedge \tau = MIN(\alpha_1',\alpha_m'). \end{cases} \quad (10)$$

The 1st formula signifies that all causes above y_i are inactive, so that the non-impeding behavior of a direct gate model is enabled.

Next, we analyze soundness of a DSDi Ψ, applying techniques in proving Lammas 1 through 4 and Theorem 3 on dual gate models. We consider $P_\Psi(e|z_1,...,z_\omega)$ defined by divorcing CPTs at e and y_i according to Eqn. (9) (similarly as DSDu). Theorem 4 establishes that $P_\Psi(e|z_1,...,z_\omega)$ is equivalent to that of the SBDi, and hence a DSDi is probabilistically equivalent to the direct gate model.

Theorem 4: Let Ψ be the DSDi of a direct NIN-AND gate model, and $P_\Psi(e|z_1,...,z_\omega)$ be defined by CPTs in Ψ. Then the CPT $P_\Psi(e|W_1,...,W_\omega)$ defined by marginalized product

$$\sum_{z_1,...,z_\omega} \left(P_\Psi(e|z_1,...,z_\omega) \prod_{i=1}^{\omega} P_\Psi(z_i|c_{i1},...,c_{i\theta_i}) \right)$$

equals $P(e|W_1,...,W_\omega)$ of the direct gate model.

VII. Trans-Causalization of NAT Models

A. Interfacing BN Segments of Gate Models

A NAT model generally consists of multiple NIN-AND gates organized into a tree. To trans-causalize a general NAT model, we create a BN segment for each gate and merge the segments into the BN segment of the NAT model, such that it encodes exactly the CPT of the NAT model. If an NIN-AND gate feeds into another in the NAT model, its effect variable is replaced with a *quasi-effect* variable.

Consider the NAT in Fig. 5 (a), where labels of causal events have been simplified (e.g., input events to gates) or omitted (e.g., output events). Suppose that the leaf gate g_2 is dual. Then g_1 is direct (since types of gates alternate by levels in a NAT). The BN segment of the NAT is shown in (b). The BN segment of g_1 consists of cause variables c_i $(i = 1,2,3)$, probabilistic auxiliary variables z_i $(i = 1,2,3)$, deterministic auxiliary variable y_1, and quasi-effect variable q. This segment is a DSDi $(\theta = 1)$, except that the variable e is renamed as q.

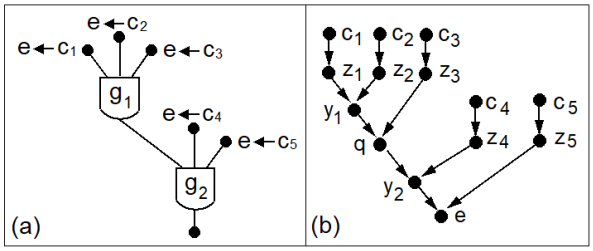

Fig. 5. (a) A NAT. (b) BN segment produced by trans-Causalization.

The BN segment of g_2 consists of cause variables c_i $(i = 4,5)$, quasi-effect variable q as an input from g_1, probabilistic auxiliary variables z_j $(j = 4,5)$, deterministic auxiliary variable y_2, and effect variable e. This segment is a DSDu $(\theta = 1)$, except that the quasi-effect variable q is treated in the same way as probabilistic auxiliary variables z_j $(j = 4,5)$.

Next, suppose that the leaf gate g_2 is direct and g_1 is dual. The BN segment of the NAT is also structured as Fig. 5 (b). However, the DSDu cannot be directly applied to dual gate

g_1 and must be modified: In the DSDu, auxiliary variables z_i and y_i, as well as the effect e, have the domain D_e. This is no longer applicable, since g_1 is not the leaf gate, and feeds into the direct gate g_2. To support non-impeding behavior of the direct gate, domains of z_i, y_i, and quasi-effect q have to be enlarged into D_a. Due to this enlargement, SC CPTs cannot be applied to z_i ($i = 1,2,3$), and MAX CPTs cannot be applied to y_1 and q.

DEDu The above situation arises whenever a dual gate feeds into a direct gate in a NAT. Hence, a new BN segment is needed. We refer to the segment as the *Divorcing Enhanced BN segment for the Dual gate model*, or ***DEDu***. Its auxiliary variables z_i and y_i have the domain D_a. Since its leaf variable is always a quasi-effect (instead of effect), we denote by q in general, and the domain of q is also D_a.

In a DEDu, auxiliary variables z_i adopt CPTs by Eqn. (6). For instance, z_i, z_2, z_3 in the example adopt SC^+ CPTs (Eqn. (7)). A new form of CPT is needed for y_1 and q. It is referred to as *PMAX*$^+$ CPTs, and encodes the following pseudo-MAX-plus (PMAX^+) function, where domains of each argument and the function range are D_a:

$$PMAX^+(\alpha_1, \alpha_2) =$$
$$\begin{cases} aaci, & if\ \alpha_i = aaci\ (i = 1,2), \\ MAX(\alpha'_1, \alpha'_m), & if\ \alpha'_1, \alpha'_m \neq aaci\ (m > 0). \end{cases}$$

The PMAX^+ CPTs at y_1 and q are the following:

$$P(\tau|\alpha_1, \alpha_2) =$$
$$\begin{cases} 1, & if\ \alpha_i = aaci\ (i = 1,2) \wedge \tau = aaci, \\ 1, & if\ \alpha'_1, \alpha'_m \neq aaci\ (m > 0) \wedge \tau = MAX(\alpha'_1, \alpha'_m). \end{cases} \quad (11)$$

The BN segment of the direct leaf gate g_2 is a DSDi ($\theta = 1$), except that the quasi-effect q is treated in the same way as auxiliary variables z_4 and z_5.

B. Enhanced BN Segments for Gate Models

We establish soundness of a DEDu Ψ, where $P_\Psi(q|z_1, ..., z_\omega)$ is defined similarly to Eqn. (9). Lemma 5 asserts the behavior of divorcing CPTs in a DEDu.

Lemma 5: Let Φ be a SBDu, $P_\Phi(e|z_1, ..., z_\omega)$ be its MAX CPT, and Ψ be the corresponding DEDu. The following CPT defined by Ψ satisfies

$$P_\Psi(q|z_1, ..., z_\omega) =$$
$$\begin{cases} 1, & if\ q = aaci\ \wedge\ \forall_i\ z_i = aaci, \\ P_\Phi(e|z_1, ..., z_\omega), & if\ \exists_i\ z_i \neq aaci. \end{cases}$$

Theorem 5 establishes probabilistic equivalence between a DEDu and the dual gate model.

Theorem 5: Let Ψ be the DEDu of a dual NIN-AND gate model, and $P_\Psi(q|z_1, ..., z_\omega)$ be defined by CPTs in Ψ. Then the CPT $P_\Psi(q|W_1, ..., W_\omega)$ defined by marginalized product

$$\sum_{z_1, ..., z_\omega} \left(P_\Psi(q|z_1, ..., z_\omega) \prod_{i=1}^{\omega} P_\Psi(z_i|c_{i1}, ..., c_{i\theta_i}) \right)$$

equals $P(e|W_1, ..., W_\omega)$ of the dual gate model when some W_i are active, and $P_\Psi(q = aaci|W_1, ..., W_\omega) = 1$ when all W_i are inactive.

DEDi Similarly to the need for DEDu when dual gates feed into direct gates, a BN segment other than DSDi is needed when a direct gate feeds into a dual gate expressed as DEDu. The domain of z_i variables in DEDu is D_a, while the domain of leaf variable in DSDi is D_e: incompatible.

We refer to the new segment as the *Divorcing Enhanced BN segment for the Direct gate model*, or ***DEDi***. Its auxiliary variables z_i and y_i have the domain D_a. Its leaf variable is always a quasi-effect (instead of effect), and we denote by q in general, and the domain of q is also D_a.

In a DEDi, z_i variables adopt CPTs of Eqn. (6), and when $\theta = 1$, SC^+ CPTs of Eqn. (7). Variables y_i and q adopt PMIN^+ CPTs of Eqn. (10).

Theorem 6 establishes probabilistic equivalence between a DEDi and the dual gate model.

Theorem 6: Let Ψ be the DEDi of a direct NIN-AND gate model, and $P_\Psi(q|z_1, ..., z_\omega)$ be defined by CPTs in Ψ. Then the CPT $P_\Psi(q|W_1, ..., W_\omega)$ defined by marginalized product

$$\sum_{z_1, ..., z_\omega} \left(P_\Psi(q|z_1, ..., z_\omega) \prod_{i=1}^{\omega} P_\Psi(z_i|c_{i1}, ..., c_{i\theta_i}) \right)$$

equals $P(e|W_1, ..., W_\omega)$ of the direct gate model when some W_i are active, and $P_\Psi(q = aaci|W_1, ..., W_\omega) = 1$ when all W_i are inactive.

C. BN Segments for NAT Models and Their Soundness

As seen from above, BN segments of NIN-AND gates need to be assigned based on their relative location in the NAT. The leaf gate of the NAT is at level 1. A gate that feeds into the leaf gate is at level 2, and so on. In Fig. 5 (a), g_2 is at level 1 and g_1 is at level 2.

In general, we set the BN segment of a gate model by Table I. It also summarizes domains of auxiliary (probabilistic and deterministic), quasi-effect, and effect variables. The quasi-effect column refers to effect at level 1.

TABLE I
SUMMARY OF VARIABLE DOMAINS

Level	Gate	Seg.	Aux. var.	Quasi-effect
1	Dual	DSDu	D_e	D_e
	Direct	DSDi	D_a	D_e
2	Dual	DEDu	D_a	D_a
	Direct	DSDi	D_a	D_e
3+	Dual	DEDu	D_a	D_a
	Direct	DEDi	D_a	D_a

Note that variable domains in DSDu (level 1) and DSDi (levels 1 and 2) are smaller ($D_e \subset D_a$). It is possible to adopt the larger domain D_a for all z_i, y_i, q, e variables in all gate BN segments. We would have only two types of BN segments, one for each type of NIN-AND gates. We have chosen to keep the variable domains as small as possible, while ensuring non-impeding behavior of NIN-AND gates. Although it leads to two extra types of BN segments and more sophisticated analysis, it ensures the best possible space efficiency and the inference efficiency (evaluated below).

The CPTs for auxiliary, quasi-effect, and effect variables in relation to type and level (L) of the gate, and its BN

segment are summarized in Table II. As shown in the above analysis, these CPTs are chosen to maintain exactness of CPTs relative to the corresponding gate model, including ensuring non-impeding behavior of NIN-AND gates downstream in the NAT.

TABLE II
SUMMARY OF VARIABLE CPTS

L	Gate	Seg.	Prob. aux.	Deter. aux.	Quasi-effect
1	Dual	DSDu	SC CPT	MAX CPT	MAX CPT
	Direct	DSDi	SC^+ CPT	$PMIN^+$ CPT	PMIN CPT
2	Dual	DEDu	SC^+ CPT	$PMAX^+$ CPT	$PMAX^+$ CPT
	Direct	DSDi	SC^+ CPT	$PMIN^+$ CPT	PMIN CPT
3+	Dual	DEDu	SC^+ CPT	$PMAX^+$ CPT	$PMAX^+$ CPT
	Direct	DEDi	SC^+ CPT	$PMIN^+$ CPT	$PMIN^+$ CPT

We show below that the BN segment of the NAT model so integrated ensures the exact $P(e|c_1,...,c_n)$ of the NAT model. We demonstrate the exactness empirically in Section IX.

Theorem 7: Let Ψ be the BN segment of a NAT model over e and $c_1,...,c_n$, integrated from DSDu, DSDi, DEDu, and DEDi of gate models with $\theta = 1$. Let $P_\Psi(e|c_1,...,c_n)$ be obtained by marginalizing out all auxiliary and quasi-effect variables from the product of all CPTs in Ψ. Then $P_\Psi(e|c_1,...,c_n)$ equals $P(e|c_1,...,c_n)$ of the NAT model.

Proof: We prove by induction on the number of levels L of the NAT. As base cases, we consider $L = 1$ and $L = 2$. When $L = 1$, the NAT has a single gate. If the gate is dual, Ψ is a DSDu, and the result follows from Theorem 3. If the gate is direct, Ψ is a DSDi, and the result follows from Theorem 4.

Suppose $L = 2$. If the leaf gate is dual, every gate at level 2 is direct. The leaf gate is assigned a DSDu (Table I), denoted by Λ. Each direct gate is assigned a DSDi (Table I). We prove the result by converting Ψ into an equivalent single DSDu.

Without losing generality, let Δ be the DSDi of a direct gate model over causes $c_1,...,c_m$ $(m < n)$, where $\theta = 1$ by assumption. Denote the leaf variable of Δ by q. As Δ feeds into Λ through quasi-effect q, we replace Δ in Ψ by making $c_1,...,c_m$ the root parents of q, and assigning q the CPT $P_\Delta(q|c_1,...,c_m)$. This is illustrated in Fig. 6, where the BN segment of NAT in Fig. 5 is reproduced in Fig. 6 (a), and replacement relative to the BN segment of g_1 is shown in (b).

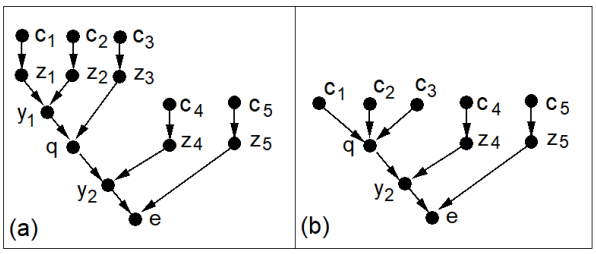

Fig. 6. (a) BN segment of NAT in Fig. 5. (b) Replacing gate BN segment.

By Theorem 4, $P_\Delta(q|c_1,...,c_m)$ equals $P(e|c_1,...,c_m)$ of the direct gate model. Hence, this replacement does not alter $P_\Psi(e|c_1,...,c_n)$. Repeating the above for each DSDi in Ψ, eventually Ψ becomes a single DSDu where $\theta > 1$, and $P_\Psi(e|c_1,...,c_n)$ is invariant. By the above argument on $L = 1$, the result follows.

Continue with $L = 2$. If the leaf gate is direct, every gate at level 2 is dual. The leaf gate is assigned a DSDi, denoted by Λ. Each dual gate is assigned a DEDu (Table I). Let Δ be the DEDu of a dual gate model over $c_1,...,c_m$ $(m < n)$, where $\theta = 1$, and its leaf variable be q. We replace Δ in Ψ by making $c_1,...,c_m$ the root parents of q, and assigning q the CPT $P_\Delta(q|c_1,...,c_m)$.

Since Λ is a DSDi, when $c_1,...,c_m$ become root parents of q, to maintain $P_\Psi(e|c_1,...,c_n)$, $P_\Lambda(q|c_1,...,c_m)$ should be specified according to Eqn. (6). By Theorem 5 on DEDu, if some cause in $c_1,...,c_m$ are active, $P_\Delta(q|c_1,...,c_m)$ equals $P(e|c_1,...,c_m)$ of the dual gate model. If all $c_1,...,c_m$ are inactive, $P_\Delta(q = aaci|c_1,...,c_m) = 1$. Hence, $P_\Delta(q|c_1,...,c_m)$ behaves exactly as specified by Eqn. (6). It follows that the above replacement does not alter $P_\Psi(e|c_1,...,c_n)$. Repeating the replacement for each DEDu in Ψ, eventually Ψ becomes a single DSDi where $\theta > 1$, and $P_\Psi(e|c_1,...,c_n)$ is invariant. By the above argument on $L = 1$, the result follows.

Assume that the theorem holds for $L \leq k$, where $k \geq 2$. We consider $L = k+1$, i.e., $L \geq 3$.

If gates at level L are dual, by Table I, each is assigned a DEDu where $\theta = 1$, and feeds into a DSDi $(L = 3)$ or DEDi $(L > 3)$. Consider one pair, where the DSDi or DEDi is denoted by Λ, and the DEDu is denoted by Δ. Let Δ feed into Λ by quasi-effect q.

We replace Δ (in Ψ) by making its cause variables, say, $c_1,...,c_m$, the root parents of q, and assigning q the CPT $P_\Delta(q|c_1,...,c_m)$. Since Λ is a DSDi or DEDi, when $c_1,...,c_m$ become root parents of q, to maintain $P_\Psi(e|c_1,...,c_n)$, $P_\Lambda(q|c_1,...,c_m)$ should be specified according to Eqn. (6). By Theorem 5 on DEDu, $P_\Delta(q|c_1,...,c_m)$ behaves exactly as specified by Eqn. (6). Hence, the replacement does not alter $P_\Psi(e|c_1,...,c_n)$. Repeating the above for each DEDu at level L in Ψ, eventually the number of levels of Ψ is reduced to k, and $P_\Psi(e|c_1,...,c_n)$ is invariant. By the inductive assumption, the result follows.

If gates at level L are direct, each is assigned a DEDi where $\theta = 1$, and feeds into a DEDu (Table I). Consider one pair, where the DEDu is denoted by Λ, the DEDi by Δ, and Δ feeds into Λ by quasi-effect q.

We replace Δ by making its cause variables, say, $c_1,...,c_m$, the root parents of q, and assigning q the CPT $P_\Delta(q|c_1,...,c_m)$. Since Λ is a DEDu, when $c_1,...,c_m$ become root parents of q, to maintain $P_\Psi(e|c_1,...,c_n)$, $P_\Lambda(q|c_1,...,c_m)$ should be specified by Eqn. (6). By Theorem 6 on DEDi, $P_\Delta(q|c_1,...,c_m)$ behaves exactly as specified by Eqn. (6). Hence, the replacement does not alter $P_\Psi(e|c_1,...,c_n)$. Repeating the above for each DEDi at level L in Ψ, eventually the number of levels of Ψ is reduced to k, while $P_\Psi(e|c_1,...,c_n)$ is invariant. By the inductive assumption, the result follows. □

VIII. TRANS-CAUSALIZATION OF NAT-MODELED BNS

Consider a NAT-Modeled BN over a set V of variables with the DAG D. Each variable with up to 2 parents is assigned a Tabular CPT, collected in a set TC. The family of each variable

with 3 or more parents is a NAT Model, collected in set NM. We denote the NAT modeled BN by $\Gamma = (V, D, TC, NM)$.

To trans-causalize Γ, for each NAT model in NM (child e plus parents $c_1, ..., c_n$), delete the link from each c_i to e in D, reconnect the family by the BN segment of the NAT model, and assign CPT to each variable in the segment (except $c_1, ..., c_n$) by Table II. Let W be the set of all auxiliary and quasi-effect variables added through the operation. Let D' be the resultant DAG over the set $V \cup W$ of nodes. Let NC be the set of all New CPTs assigned by Table II. We denote the trans-causalized BN by $\Omega = (V \cup W, D', TC, NC)$.

Consider an example NAT-modeled BN Γ in Fig. 7, where the NAT model over family of v_8 is shown with simplified labeling, and all variables are ternary. The gate g_3 is direct, and g_1 and g_2 are dual.

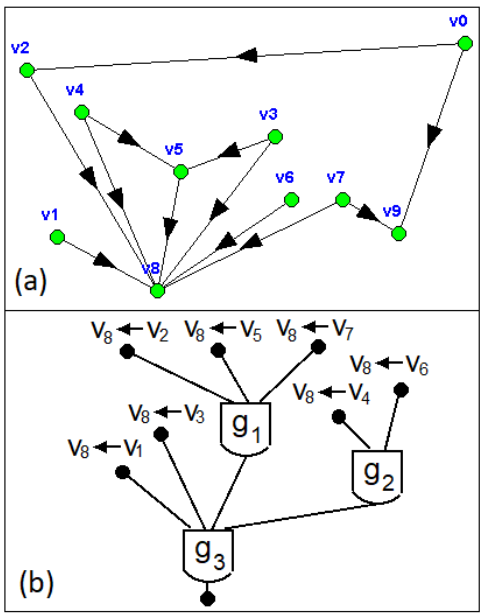

Fig. 7. (a) DAG of NAT-model BN. (b) NAT-model over family of v_8.

The trans-causalized BN Ω from Γ is shown in Fig. 8. For causes v_i $(i = 1, ..., 7)$ in that order, the probabilistic auxiliary variables are x_{10}, x_{16}, x_{11}, x_{20}, x_{17}, x_{21}, x_{18}, respectively. For gate g_2, the quasi-effect is q_{13}. For gate g_1, the deterministic auxiliary variable is y_{19} and the quasi-effect is q_{12}. For gate g_3, the deterministic auxiliary variables are y_{14} and y_{15}.

Suppose that all variables in Γ are ternary. If for each NAT model in Γ, variable e is assigned an equivalent tabular CPT, the resultant BN has 6642 numerical parameters (values in all CPTs). On the other hand, Ω has 489 parameters.

Theorem 8 establishes that the trans-causalized BN is an exact representation of the NAT modeled BN.

Theorem 8: Let $\Gamma = (V, D, TC, NM)$ be a NAT modeled BN, and $\Omega = (V \cup W, D', TC, NC)$ be the trans-causalized BN from Γ. Let $P_\Gamma(V)$ be the joint probability distribution (JPD) of Γ, and $P_\Omega(V, W)$ be the JPD of Ω. Then

$$\sum_{w \in W} P_\Omega(V, W) = P_\Gamma(V).$$

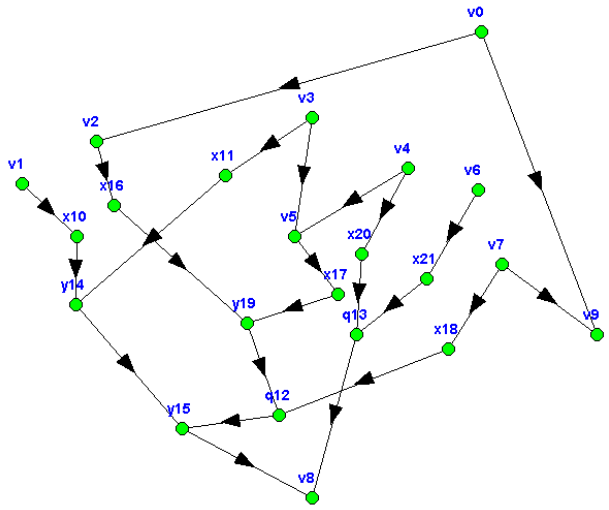

Fig. 8. Trans-causalization of NAT-modeled BN in Fig. 7 (a).

Theorem 8 shows that trans-causalization of a NAT-modeled BN is a systematic way to introduce hidden variables, which improves space efficiency, eliminates representational heterogeneity in NAT-modeled BNs, while preserving the JPD over variables of the BN. Therefore, probabilistic inference with trans-causalized BNs can be performed using any standard inference algorithm. As long as only observations over variables in V are entered (as W is made of hidden variables that are unobservable), any posteriors over unobserved variables in V are exact. In Section IX, we demonstrate that posterior marginals thus computed are exact as computed from the original NAT-modeled BNs.

We consider space complexity of trans-causalized BNs. Let $\Gamma = (V, D, TC, NM)$ be a NAT modeled BN, and Ω be the trans-causalized BN from Γ. Denote the number of variables by $N = |V|$, the largest domain size of variables by κ, and the largest number of parents per variable by n.

The number of NAT models in Γ is $O(N)$. After trans-causalization with divorcing, a variable of n parents is replaced by $n-1$ variables each of two parents and n variables each of one parent. For example, the BN in Fig. 7 (a) has a single NAT model over the family of v_8 with 7 parents. Hence, v_8 is replaced by $6+7$ variables of smaller families, growing the total number of variables in the BN from 10 to $10-1+13=22$ in Fig. 8. The total size of all CPTs in the BN segment is $O((n-1)\ \kappa^3 + n\ \kappa^2) = O(n\ \kappa^3)$. The space complexity of Ω is then $O(N\ n\ \kappa^3)$. In comparison, if a BN is obtained from Γ by converting each NAT model into a tabular CPT, its space is $O(N\ \kappa^n)$. For $N = 100$, $\kappa = 4$, and $n = 10$, we have $N\ n\ \kappa^3 = 64,000$, and $N\ \kappa^n = 104,857,600$.

IX. Experiments

We evaluated trans-causalization empirically through several experiments. Each experiment is conducted on an extensive collection of NAT models or NAT-modeled BNs. To keep this section concise, we report results from a subset of cases

evaluated in each experiment, which are representative of the corresponding collection.

The 1st experiment compares space of a NAT model CPT expressed as table, against total spaces of all CPTs when it is trans-causalized with and without parent divorcing. The spaces are measured by the number of numerical parameters, and are labeled as TAB, TPD (with), and TRC (without), respectively. TAB refers to TABular, TRC refers to TRans-Causal, and TPD refers to Trans-causalization with Parent Divorcing. We report results when numbers of causes per NAT model are $n = 5, 11$, and uniform domain sizes of causes and effect per NAT model are $\kappa = 5, 7$, with 30 random NAT topologies per combination of (n, κ). These amount to evaluation of $2 * 2 * 30 = 120$ distinctly structured NAT models.

Figs. 9 and 10 show spaces in $log10$. TAB space is completely determined by (n, κ), and is constant. TRC space is sensitive to NAT topology, and is often less than TAB space. This demonstrates the effectiveness of trans-causalization. For some NATs, TRC space is slightly more than TAB space. For instance, the 11th TRC space for $(n = 11, \kappa = 7)$ is more than the TAB space, whose NAT has two gates and one of them has 10 inputs. TPD space is only slightly sensitive to NAT topology and further improves upon TRC space. In Section X, we discuss further the relation between trans-causalization and parent divorcing.

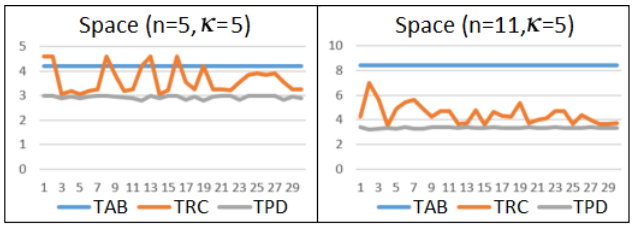

Fig. 9. TAB, TRC, and TPD spaces with $\kappa = 5$.

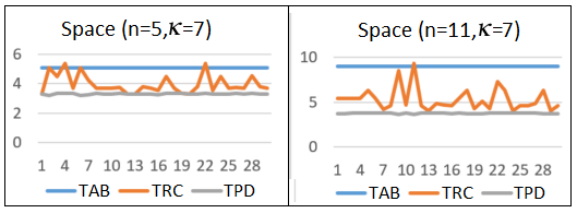

Fig. 10. TAB, TRC, and TPD spaces with $\kappa = 7$.

TPD space is always the most efficient among the three. When κ is fixed, TPD space becomes more efficient than TAB as n grows (compare their difference in Fig. 10 in the left with that in the right). The same trend holds when n is fixed and κ grows. This can be seen by comparing TAB-TPD difference in Fig. 9 (left) with that in Fig. 10 (left). On average, the TAB/TPD ratio for $(n = 5, \kappa = 7)$ is 3.19 times as high as that for $(n = 5, \kappa = 5)$. For $n = 11$ (Fig. 9 (right) versus Fig. 10 (right)), on average, the TAB/TPD ratio for $(n = 11, \kappa = 7)$ is 1.62 times as high as that for $(n = 11, \kappa = 5)$. Hence, TPD becomes more advantageous as (n, κ) scale up. For $(n = 11, \kappa = 7)$, TPD space is 5 orders of magnitude more efficient than TAB space. In the remaining experiments, our implementation of trans-causalization is always enhanced with parent divorcing.

The 2nd experiment evaluates the impact of trans-causalization on inference efficiency, where the inference method is LP. We simulated NAT-modeled BNs with 100 variables per BN. We report results where the maximum number of parents per variable in each BN is bounded at $m = 10, 12$, respectively. The uniform domain size of all variables is controlled as $\kappa = 2, 3$, respectively. The structural density of BNs is controlled by adding $w = 5, 15, 25, 35, 45, 55\%$ of links to a singly connected network, respectively. For each (m, κ, w) combination, we simulated 10 BNs. This amounts to $2 \times 2 \times 6 = 24$ distinct (m, κ, w) combinations and 240 NAT-modeled BNs. We limit the structural density to $w = 55$ as inference times of the two alternative methods (see below) tend to converge beyond $w = 55$.

For each NAT-modeled BN, we created a *normalized* BN (NM-BN) where each NAT model is expanded into a tabular CPT, and a trans-causalized BN (TC-BN). Both NM-BN and TC-BN are compiled for LP, conditioned on the same observation over 10% of randomly selected variables. For each pair of NM-BN and TC-BN, LP resulted in identical posterior marginals, which empirically verifies exactness of trans-causalization. LP runtimes using a desktop of 3.4 GHz clock speed, are summarized in Fig. 11, where legends are identical.

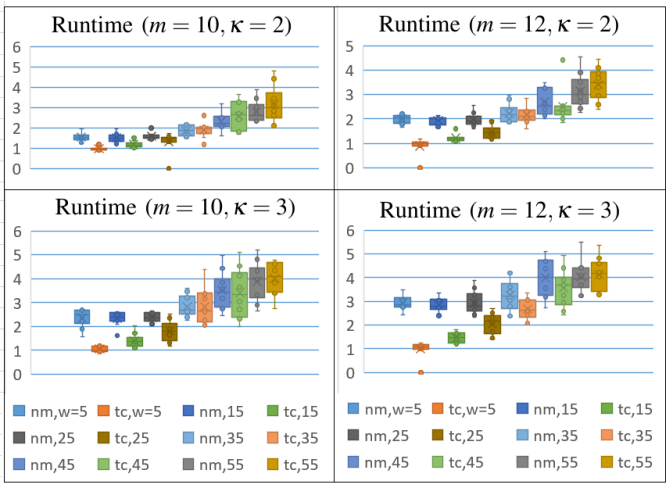

Fig. 11. LP runtimes (msec in log10) for NM-BNs and TC-BNs.

The inference computation becomes more expensive as m, κ and w grow. For sparse BN structures, as inference becomes harder, TC-BNs become more advantageous than NM-BNs. For instance, with $w = 5$, as m and κ grow, the runtime by TC-BNs become significantly less than NM-BNs. At $(m = 12, \kappa = 3, w = 5)$, LPs with TC-BNs are two orders of magnitude fasters than NM-BNs.

Furthermore, as m and κ grow, the range of structural densities where TC-BNs are more efficient than NM-BNs grows as well. For instance, for $(m = 10, \kappa = 3)$, TC-BNs and NM-BNs tie in runtime around $w = 45$. As m grows to 12, the corresponding structural density grows to $w = 55$.

The 3rd experiment evaluates the impact of trans-causalization on inference efficiency, where the inference is performed through SPNs [17], [18]. Each simulated NAT-modeled BN has 100 variables. We report results where the maximum number of parents per variable m, the uniform domain size of all variables κ, and the structural density w are controlled as $m = 6,8,10,12$, $\kappa = 2,3$, and $w = 5,10,15,20,25,30\%$, respectively. For each (m,κ,w) combination, we simulated 10 BNs. This amounts to $4\times 2\times 6 = 48$ distinct (m,κ,w) combinations and 480 NAT-modeled BNs.

Each NAT-modeled BN is converted into a trans-causalized BN (TC-BN), and a multiplicatively factorized NAT-modeled BN (MF-BN) [14]. The TC-BN is compiled for inference by LP, and also compiled into a SPN for inference. The MF-BN is only compiled for LP. Hence, each NAT-modeled BN is subject to 3 alternative combinations of runtime representation and inference method, all conditioned on the same observation over 10% of randomly selected variables. We refer to corresponding runtimes as TC-LP, TC-SPN, and MF-LP, respectively. We control the structure density at $w = 30$ due to an exponential growth of SPN compilation time. For instance, other parameters being identical, when $w = 35$, compilation of the BN into SPN often took over two hours, whereas the compilation times for TC-LP and MF-LP took at most a minute.

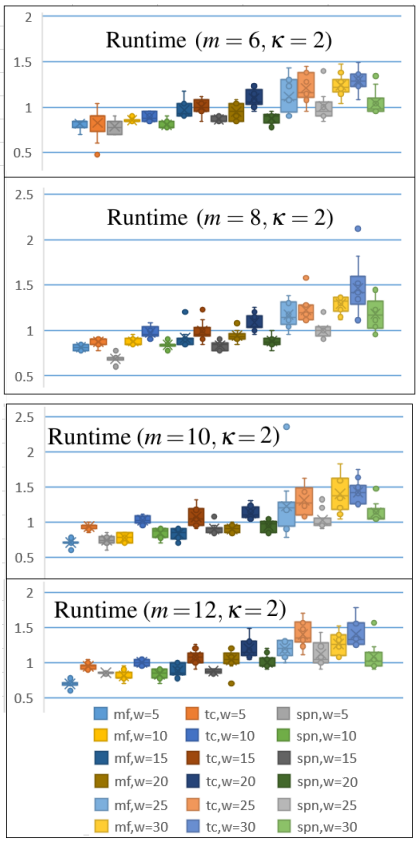

Fig. 12. Runtimes of MF-LP (mf), TC-LP (tc), and TC-SPN (spn), where $\kappa = 2$.

Figs. 12 and 13 show runtimes (msec) in $log10$. For sparse BN structures (e.g., $w = 5,10$), MF-LP and TC-SPN runtimes are comparable, but each is generally less than TC-LP runtime.

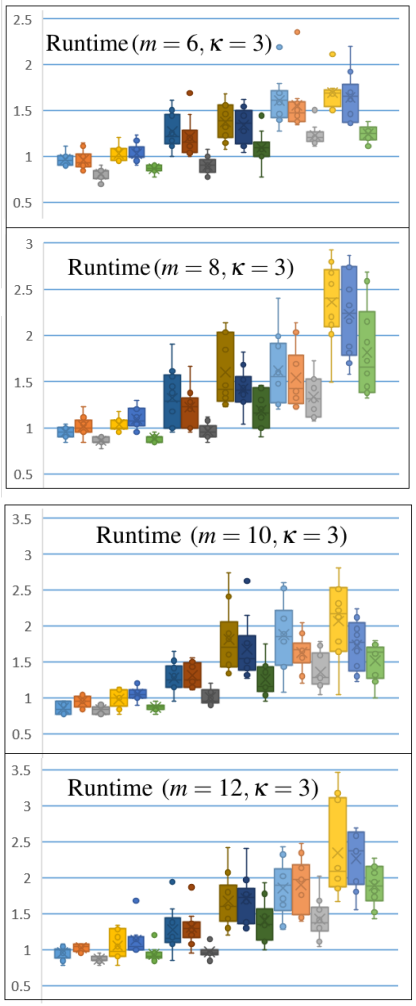

Fig. 13. Runtimes of MF-LP, TC-LP, and TC-SPN, where $\kappa = 3$.

As BN structures become denser, TC-SPN inference becomes more efficient than MF-LP and TC-LP inference. Furthermore, as κ grows from 2 to 3, variability in runtime grows for all methods, and TC-SPN inference gains further advantage against the competing methods. At $(m = 12, \kappa = 3, w = 30)$, the upper runtime of TC-SPN inference is an order of magnitude less than that of TC-LP inference, and two orders of magnitude less than that of MF-LP inference.

As the focus of the experiment is to compare performance of trans-causalized BNs with competitive alternatives, we did not include MF-SPN and NM-BN in the experiment, which should not affect the results. For readers interested in relative performance of MF-SPN and MF-LP, it is expected that MF-SPN will be more efficient than MF-LP. The reason is that potentials introduced during multiplicative factorization of NAT-modeled BNs contain a sufficient amount of extreme probability values, and can be explored by MF-SPN. For comparison between TC-SPN and NM-BN by LP, from the superior performance of TC-SPN over TC-LP, and that of TC-BN over NM-BN in the 2nd experiment, it follows that TC-SPN expects a superior performance over NM-BN by LP. For NM-BN by SPN, a performance similar to NM-BN by LP is

expected, since NM-BN does not contain a sufficient amount of extreme probability values.

The 4th experiment compares efficiency of inference with trans-causalized BNs and that with multiplicatively factorized NAT-modeled BNs. Since the space complexity for the later is exponential on the domain size of the effect variable [14], the main parameter we control in the experiment is domain size κ. NAT-modeled BNs are simulated with 100 variables each. We fix $m = 8, w = 15$ and vary κ in the range $\kappa = 2, 4, 6$. For each (m, κ, w) combination, we simulated 10 BNs. This amounts to a total of 30 BNs.

Each NAT-modeled BN is converted into a trans-causalized BN (TC-BN), and a multiplicatively factorized NAT-modeled BN (MF-BN). The TC-BN is compiled into a SPN for inference, and the MF-BN is compiled for LP, which we refer to below as TC-SPN inference and MF-LP inference. For each NAT-modeled BN, the 2 alternative inference methods were run, conditioned on the same observation over 10% of randomly selected variables. This amounts to a total of 60 inference runs.

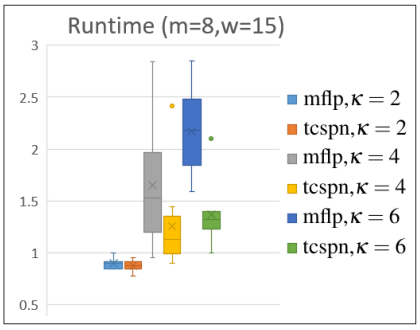

Fig. 14. Runtimes of MF-LP and TC-SPN, where $m = 8, w = 15$.

Fig. 14 shows runtimes (msec) in $log10$. At $\kappa = 2$, runtimes are similar between MF-LP and TC-SPN. As κ increases to $\kappa = 4, 6$, the advantage of TC-SPN over MF-LP becomes clear. At $\kappa = 6$, TC-SPN is on average 7.2 times as fast as MF-LP. This demonstrates that trans-causalizing NAT-modeled BNs is superior than multiplicatively factorizing NAT-modeled BNs when effect variables have larger domains. Since the space complexity of TC-BN is polynomial (Section VIII) while that of MF-BN is exponential on the domain size of the effect variable [14], the result confirms that the qualitative difference extends to the time complexity of inference.

Although we did not include MF-SPN in this experiment, the polynomial versus exponential space complexity between TC-BN and MF-BN expects a weaker performance of MF-SPN relative to TC-SPN, at least during SPN compilation, and likely also during inference.

X. Discussion and Conclusion

A. Discussion on Representational Issues

We discuss issues on how NAT models relate to other CIMs which are of interest to readers. First, we consider the representation of uncertain causal relations. At least 3 alternative approaches can be identified. Pearl [1] presents noisy-OR with deterministic causes whose activeness guarantees activation of the effect. The uncertainty lies in the inhibitor associated with each cause. The inhibitor is active probabilistically and when it is active, the causation is blocked. A second approach is to introduce auxiliary variables one per cause as the intermediate effect, and to combine them into the final effect, e.g., [22]. The third approach is to model causes directly as being uncertain, which render the effect active probabilistically, and to use causal events as the basis of representation. This is the approach taken by recursive noisy-OR [7] and the NAT models.

There are often alternative formalisms for representing a given entity. For instance, factorized probabilistic knowledge can be represented by BNs, Markov networks, chain graphs, among others. Even BNs can be encoded through DAG structures, junction tree structures, or sum-product networks. Each of the formalism has its advantages and limitations. This is the same regarding representational approaches of CIMs:

An essential strength of NAT models is to be able to express both reinforcing and undermining causal interactions, as well as their recursive mixture, thus generalizing the commonly applied noisy-OR, noisy-MAX, and DeMorgan. Using uncertain causes, causal success, and causal failure as the basis of representation are not only intuitive, but also transparent with respect to the reinforcing and undermining causal interactions. For instance, in Fig. 7 (b), it is immediately clear that V_2 and V_6 are undermining each other since their common gate is g_3 which is direct. Similarly, V_2 and V_7 are reinforcing since their common gate is g_1 which is dual. Furthermore, no inhibitors nor auxiliary variables are needed, making the NAT model representation arguably easier to comprehend. However, the NAT models are not directly compatible with the BN structures, which motivates this work on trans-causalization.

On the other hand, the approach based on auxiliary variables (2nd approach above) is compatible with the normal BN structure and d-separation. As can be seen, trans-causalization is based on introducing such auxiliary variables. Therefore, trans-causalization can be viewed as bridging the 2nd and 3rd approaches in NAT model representation. However, although trans-causalization makes a NAT-modeled BN compatible with a normal BN, some representational advantages of native NAT models are lost. For instance, it is impossible to use Fig. 8 to decide the causal interaction between V_2 and V_6. More importantly, the semantics of causal interactions by reinforcing and undermining will be completely lost from the trans-causalized structure. It is also unintuitive that auxiliary and quasi-effect variables sometimes have the same domain as the effect, but other times must add the *aaci* value to the domain (Table I).

As our analysis shows, NAT-modeled BNs are equivalent to their trans-causalization. One may ask whether NAT-models should be defined in the trans-causalized version. As the native representation, we believe that intuitiveness and comprehensibility are more important, and NAT-models are defined in a way that serves these purposes well.

We presented trans-causalization and its enhancement by divorcing. An interesting question arises whether divorcing should be an essential element of trans-causalization rather

than an enhancement. The 1st experiment in Section IX shed some light. The result clearly demonstrates that trans-causalization, as we defined, works on its own, and divorcing is indeed an enhancement.

B. Conclusion

The main contribution of this work is the novel trans-causalization framework, by which a NAT-modeled BN is converted into a trans-causalized BN for inference computation. We formally establish that trans-causalization of BNs is exact. While the space complexity of BNs with tabular CPTs is exponential on the number of causes per effect variable (BN family sizes), and the space complexity of multiplicatively factorized NAT-modeled BNs is exponential on the domain size of effect variables, the space complexity of trans-causalized BNs is polynomial.

Trans-causalization gains the space efficiency by causal independence between causes of the same NIN-AND gate, as well as causal independence between causes of different gates in the same NAT. The space efficiency is further enhanced with divorcing by exploiting the deterministic CPTs at quasi-effect variables.

The above formal contributions are further enhanced by several empirical results. First, the superior space property of trans-causalized BNs is demonstrated as both BN family sizes and variable domain sizes scale up. Second, we demonstrated that inference with trans-causalized BNs is significantly more efficient than regular BNs (up to 2 orders of magnitude faster) for a range of sparse structural densities. Third, our comparison of TC-SPN, TC-LP, and MF-LP inference showed that, for denser BN structures and larger variable domains, SPN inference with trans-causalized BNs is more efficient than LP based inference, which is in turn more efficient than MF-LP inference. Finally, we confirmed empirically that the superior space complexity of trans-causalized BNs (polynomial) over that of multiplicatively factorized NAT-modeled BNs directly extend to their time complexities in inference, as variable domain sizes scale up.

The contributions of this work have several implications. They provide evidence that NAT-modeled BNs form a class of sufficiently expressive while efficient probabilistic graphical models (of polynomial space). For a range of sparse structures, the space efficiency of NAT-modeled BNs translates to significantly improved inference efficiency. As probabilistic reasoning in general BNs are NP-hard, e.g., [23], NAT-modeled BNs form a promising subclass of BNs for tractable inference, which may serve as the target representation for either knowledge acquisition or machine learning.

ACKNOWLEDGEMENT

Financial support from NSERC Discovery Grant to first author is acknowledged.

REFERENCES

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

[2] S. Li, T. Tryfonas, G. Russell, and P. Andriotis, "Risk assessment for mobile systems through a multilayered hierarchical Bayesian network," *IEEE Trans. Cybernetics*, vol. 46, no. 8, pp. 1749–1759, 2016.

[3] M. Henrion, "Some practical issues in constructing belief networks," in *Uncertainty in Artificial Intelligence 3*, L. Kanal, T. Levitt, and J. Lemmer, Eds. Elsevier Science Publishers, 1989, pp. 161–173.

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

[5] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller, "Context-specific independence in Bayesian networks," in *Proc. 12th Conf. on Uncertainty in Artificial Intelligence*, 1996, pp. 115–123.

[6] S. Galan and F. Diez, "Modeling dynamic causal interaction with Bayesian networks: temporal noisy gates," in *Proc. 2nd Inter. Workshop on Causal Networks*, 2000, pp. 1–5.

[7] J. Lemmer and D. Gossink, "Recursive noisy OR - a rule for estimating complex probabilistic interactions," *IEEE Trans. on System, Man and Cybernetics, Part B*, vol. 34, no. 6, pp. 2252–2261, Dec 2004.

[8] Y. Xiang, "Non-impeding noisy-AND tree causal models over multi-valued variables," *International J. Approximate Reasoning*, vol. 53, no. 7, pp. 988–1002, Oct 2012.

[9] Y. Xiang and M. Truong, "Acquisition of causal models for local distributions in Bayesian networks," *IEEE Trans. Cybernetics*, vol. 44, no. 9, pp. 1591–1604, 2014.

[10] P. Maaskant and M. Druzdzel, "An independence of causal interactions model for opposing influences," in *Proc. 4th European Workshop on Probabilistic Graphical Models*, M. Jaeger and T. Nielsen, Eds., Hirtshals, Denmark, 2008, pp. 185–192.

[11] J. Vomlel and P. Tichavsky, "An approximate tensor-based inference method applied to the game of Minesweeper," in *Proc. 7th European Workshop on Probabilistic Graphical Models, Springer LNAI 8745*, 2012, pp. 535–550.

[12] S. Woudenberg, L. van der Gaag, and C. Rademaker, "An intercausal cancellation model for Bayesian-network engineering," *Inter. J. Approximate Reasoning*, vol. 63, pp. 32–47, 2015.

[13] Y. Xiang and Q. Jiang, "NAT model based compression of Bayesian network CPTs over multi-valued variables," *Computational Intelligence*, vol. 34, no. 1, pp. 219–240, 2018.

[14] Y. Xiang and Y. Jin, "Efficient probabilistic inference in Bayesian networks with multi-valued NIN-AND tree local models," *Int. J. Approximate Reasoning*, vol. 87, pp. 67–89, 2017.

[15] A. Madsen and F. Jensen, "Lazy propagation: A junction tree inference algorithm based on lazy evaluation," *Artificial Intelligence*, vol. 113, no. 1-2, pp. 203–245, 1999.

[16] Y. Xiang and B. Baird, "Compressing Bayesian networks: Swarm-based descent, efficiency, and posterior accuracy," in *Canadian AI 2018, LNAI 10832*, E. Bagheri and J. Cheung, Eds. Springer, 2018, pp. 3–16.

[17] A. Darwiche, "A differential approach to inference in Bayesian networks," *J. ACM*, vol. 50, no. 3, pp. 280–305, 2003.

[18] H. Poon and P. Domingos, "Sum-product networks: A new deep architecture," in *Proc. 12th Conf. on Uncertainty in Artificial Intelligence*, 2011, pp. 2551–2558.

[19] H. Zhao, M. Melibari, and P. Poupart, "On the relationship between sum-product networks and Bayesian networks," in *Proc. 32nd Inter. Conf. Machine Learning*, 2015, pp. 116–124.

[20] Y. Xiang, "Acquisition and computation issues with NIN-AND tree models," in *Proc. 5th European Workshop on Probabilistic Graphical Models*, P. Myllymaki, T. Roos, and T. Jaakkola, Eds., Finland, 2010, pp. 281–289.

[21] K. Olesen, U. Kjrulff, F. Jensen, F. Jensen, B. Falck, S. Andreassen, and S. Andersen, "A munin network for the median nerve-a case study on loops," *Applied Artificial Intelligence*, vol. 3, no. 2-3, pp. 385–403, 1989.

[22] D. Heckerman and J. Breese, "Causal independence for probabilistic assessment and inference using Bayesian networks," *IEEE Trans. on System, Man and Cybernetics*, vol. 26, no. 6, pp. 826–831, Nov 1996.

[23] G. Cooper, "The computational complexity of probabilistic inference using Bayesian belief networks," *Artificial Intelligence*, vol. 42, no. 2-3, pp. 393–405, 1990.

Yang Xiang Yang Xiang is a Professor of Computer Science at the University of Guelph, Canada. He received his Ph.D. from the University of British Columbia, Canada, in 1992. His research interests include probabilistic and decision-theoretic graphical models, multiagent graphical models, collaborative design, decision making and planing, distributed constraint satisfaction, knowledge discovery and data mining, diagnosis and scheduling. He authored a monograph on multiagent probabilistic reasoning using graphical models, published more than 130 artitcals, and developed WebWeavr, a comprehensive Java toolkit for decision support with graphical models.

Dylan Loker Dylan Loker received his Bachelors and Masters degrees in Computer Science from the University of Guelph, Canada, in 2016 and 2018, respectively. His research focus was on causal models in artificial intelligence. Since graduating he has worked for Brock Solutions Inc. and Computational Hydraulics International. He received the Best Paper award in 2018 Canadian AI Conference for a paper co-authored with Y. Xiang.

SUPPLIMENTARY MATERIALS

Proof of Theorem 1

To prove that Φ and the gate model have the identical $P(e|W_1,...,W_\omega)$, we show that they have the identical cumulative distribution. Without losing generality, we assume that all causes in $W_1,...,W_m$ are active and those in $W_{m+1},...,W_\omega$ are inactive, where $m \leq \omega$. The dual gate model is characterized by Eqn. (2), reproduced below:

$$P(e < e^k \leftarrow \underline{w}_1^+,...,\underline{w}_m^+) = \prod_{i=1}^{m} P(e < e^k \leftarrow \underline{w}_i^+), \quad (k = 1,...,\eta).$$

Since $e < e^k$ is equivalent to $e \leq e^{k-1}$, we rewrite the above as

$$P(e \leq e^k \leftarrow \underline{w}_1^+,...,\underline{w}_m^+) = \prod_{i=1}^{m} P(e \leq e^k \leftarrow \underline{w}_i^+), (k = 0,...,\eta-1),$$

which is a cumulative causal distribution. If Φ has the identical cumulative distribution (shown below), then $P_\Phi(e|W_1,...,W_\omega)$ is also identical to $P(e|W_1,...,W_\omega)$ of the dual gate model.

Since only $W_1,...,W_m$ are active, and the CPT by Eqn. (5) encodes the *MAX* function, we have

$$P_\Phi(e \leq e^k \leftarrow \underline{w}_1^+,...,\underline{w}_m^+)$$

$$= \sum_{MAX(z_1,...,z_\omega) \leq e^k} P_\Phi(z_1,...,z_\omega|\underline{w}_1^+,...,\underline{w}_m^+,\underline{w}_{m+1}^0,...,\underline{w}_\omega^0).$$

Since $MAX(z_1,...,z_\omega) \leq e^k$ iff $z_i \leq e^k$ for $i = 1,...,\omega$, the above is equal to

$$\sum_{z_1 \leq e^k,...,z_\omega \leq e^k} P_\Phi(z_1,...,z_\omega|\underline{w}_1^+,...,\underline{w}_m^+,\underline{w}_{m+1}^0,...,\underline{w}_\omega^0)$$

$$= \sum_{z_1 \leq e^k} ... \sum_{z_\omega \leq e^k} P_\Phi(z_1,...,z_\omega|\underline{w}_1^+,...,\underline{w}_m^+,\underline{w}_{m+1}^0,...,\underline{w}_\omega^0).$$

By the DAG structure of Φ, z_i is independent of z_j for $i \neq j$ given W_i. Hence, the above equals

$$\sum_{z_1 \leq e^k} ... \sum_{z_\omega \leq e^k} P_\Phi(z_1|\underline{w}_1^+) \ ... \ P_\Phi(z_m|\underline{w}_m^+)$$

$$P_\Phi(z_{m+1}|\underline{w}_{m+1}^0) \ ... \ P_\Phi(z_\omega|\underline{w}_\omega^0)$$

$$= \sum_{z_1 \leq e^k} P_\Phi(z_1|\underline{w}_1^+) \ ... \ \sum_{z_m \leq e^k} P_\Phi(z_m|\underline{w}_m^+)$$

$$\sum_{z_{m+1} \leq e^k} P_\Phi(z_{m+1}|\underline{w}_{m+1}^0) \ ... \ \sum_{z_\omega \leq e^k} P_\Phi(z_\omega|\underline{w}_\omega^0).$$

Since $\sum_{z_i \leq e^k} P_\Phi(z_i|\underline{w}_i^0) = 1$ for $i = m+1,...,\omega$, the above equals

$$\sum_{z_1 \leq e^k} P_\Phi(z_1|\underline{w}_1^+) \ ... \ \sum_{z_m \leq e^k} P_\Phi(z_m|\underline{w}_m^+) = \prod_{i=1}^{m} P_\Phi(z_i \leq e^k \leftarrow \underline{w}_i^+).$$

From Eqn. (3), the above equals $\prod_{i=1}^{m} P(e \leq e^k \leftarrow \underline{w}_i^+)$. Hence, we have

$$P_\Phi(e \leq e^k \leftarrow \underline{w}_1^+,...,\underline{w}_m^+) = \prod_{i=1}^{m} P(e \leq e^k \leftarrow \underline{w}_i^+). \qquad \square$$

Proof of Theorem 2

We show that Φ and the gate model have the identical cumulative distribution, and hence the identical $P(e|W_1,...,W_\omega)$. Without losing generality, we assume that all causes in $W_1,...,W_m$ are active ($m \leq \omega$) and those in $W_{m+1},...,W_\omega$ are inactive. The direct gate model is characterized by Eqn. (1):

$$P(e \geq e^k \leftarrow \underline{w}_1^+,...,\underline{w}_m^+) = \prod_{i=1}^{m} P(e \geq e^k \leftarrow \underline{w}_i^+), \quad (k = 0,...,\eta).$$

It is a cumulative distribution because, for $k = 1,...,\eta$, we have

$$P(e \geq e^{k-1} \leftarrow \underline{w}_1^+,...,\underline{w}_m^+) =$$

$$P(e = e^{k-1} \leftarrow \underline{w}_1^+,...,\underline{w}_m^+) + P(e \geq e^k \leftarrow \underline{w}_1^+,...,\underline{w}_m^+),$$

and in addition $P(e \geq e^0 \leftarrow \underline{w}_1^+,...,\underline{w}_m^+) = 1$. If Φ has the identical cumulative distribution (shown below), then $P_\Phi(e|W_1,...,W_\omega)$ is also identical to $P(e|W_1,...,W_\omega)$ of the direct gate model.

Since only $W_1,...,W_m$ are active and, when they do, the CPT by Eqn. (8) encodes the *MIN* function, we have

$$P_\Phi(e \geq e^k \leftarrow \underline{w}_1^+,...,\underline{w}_m^+) =$$

$$\sum_{MIN(z_1,...,z_\omega) \geq e^k} P_\Phi(z_1,...,z_\omega|\underline{w}_1^+,...,\underline{w}_m^+,\underline{w}_{m+1}^0,...,\underline{w}_\omega^0).$$

Since $MIN(z_1,...,z_\omega) \geq e^k$ iff $z_i \geq e^k$ for $i = 1,...,\omega$, the above equals

$$\sum_{z_1 \geq e^k,...,z_\omega \geq e^k} P_\Phi(z_1,...,z_\omega|\underline{w}_1^+,...,\underline{w}_m^+,\underline{w}_{m+1}^0,...,\underline{w}_\omega^0)$$

$$= \sum_{z_1 \geq e^k} ... \sum_{z_\omega \geq e^k} P_\Phi(z_1,...,z_\omega|\underline{w}_1^+,...,\underline{w}_m^+,\underline{w}_{m+1}^0,...,\underline{w}_\omega^0).$$

By the DAG structure in Φ, z_i is independent of z_j for $i \neq j$ given W_i. Hence, the above equals

$$\sum_{z_1 \geq e^k} ... \sum_{z_\omega \geq e^k} P_\Phi(z_1|\underline{w}_1^+) \ ... \ P_\Phi(z_m|\underline{w}_m^+)$$

$$P_\Phi(z_{m+1}|\underline{w}_{m+1}^0) \ ... \ P_\Phi(z_\omega|\underline{w}_\omega^0)$$

$$= \sum_{z_1 \geq e^k} P_\Phi(z_1|\underline{w}_1^+) \ ... \ \sum_{z_m \geq e^k} P_\Phi(z_m|\underline{w}_m^+)$$

$$\sum_{z_{m+1} \geq e^k} P_\Phi(z_{m+1}|\underline{w}_{m+1}^0) \ ... \ \sum_{z_\omega \geq e^k} P_\Phi(z_\omega|\underline{w}_\omega^0).$$

By the 1st formula of Eqn. (6) and relation $e^\eta < aaci$, we have $\sum_{z_i \geq e^k} P_\Phi(z_i|\underline{w}_i^0) = 1$ for $i = m+1,...,\omega$. Hence, the above equals

$$\sum_{z_1 \geq e^k} P_\Phi(z_1|\underline{w}_1^+) \ ... \ \sum_{z_m \geq e^k} P_\Phi(z_m|\underline{w}_m^+) = \prod_{i=1}^{m} P_\Phi(z_i \geq e^k \leftarrow \underline{w}_i^+).$$

From the 2nd formula of Eqn. (6), the above is equal to $\prod_{i=1}^{m} P(e \geq e^k \leftarrow \underline{w}_i^+)$. Hence, we have

$$P_\Phi(e \geq e^k \leftarrow \underline{w}_1^+,...,\underline{w}_m^+) = \prod_{i=1}^{m} P(e \geq e^k \leftarrow \underline{w}_i^+). \qquad \square$$

Proof of Lemma 1

Proof: [Sufficiency] We prove by induction on ω. When $\omega = 2$, the only factor $P_\Psi(e|z_1,z_2) = 1$ by Eqn. (5). Assume that the product has value 1 when $\omega = k \geq 2$.

Consider the case $\omega = k+1$, where the number of factors is $\omega - 1 = k$. Denote product of the first $k-1$ factors by T, and the overall product by $T * P_\Psi(e|y_{\omega-2}, z_\omega)$. Since domains of y_i and e are identical, factors of T are exactly the $k-1$ factors for the case $\omega = k$. By inductive assumption, we have $T = 1$. Since $y_{\omega-2} = MAX(z_1,...,z_{\omega-1})$ and $e = MAX(z_1,...,z_\omega)$, we have $e = MAX(y_{\omega-2}, z_\omega)$. Hence, $P_\Psi(e|y_{\omega-2}, z_\omega) = 1$ by Eqn. (5). Therefore, $T * P_\Psi(e|y_{\omega-2}, z_\omega) = 1$.

[Necessity] Assume that the product of $\omega - 1$ factors has value 1. We prove by induction on ω that the $\omega - 1$ MAX conditions hold. When $\omega = 2$, the product is $P_\Psi(e|z_1,z_2) = 1$. By Eqn. (5), $e = MAX(z_1,z_2)$ holds.

Assume that the MAX conditions hold when $\omega = k \geq 2$. Consider the case $\omega = k+1$, where the number of factors is $\omega - 1 = k$. Denote product of the first $k-1$ factors by T, and the overall product by $T * P_\Psi(e|y_{\omega-2}, z_\omega)$. Since domains of y_i and e are identical, factors of T are exactly the $k-1$ factors for the case $\omega = k$. Since $P_\Psi(e|y_{\omega-2}, z_\omega) \leq 1$, $T * P_\Psi(e|y_{\omega-2}, z_\omega) = 1$ implies $T = 1$ and $P_\Psi(e|y_{\omega-2}, z_\omega) = 1$.

From $T = 1$ and inductive assumption, the $k - 1 = \omega - 2$ MAX conditions below hold:

$$y_1 = MAX(z_1,z_2),...,y_{\omega-2} = MAX(z_1,...,z_{\omega-1}).$$

From $P_\Psi(e|y_{\omega-2}, z_\omega) = 1$ and Eqn. (5), the $\omega - 1$th MAX condition $e = MAX(z_1,...,z_\omega)$ holds. □

Proof of Lemma 2

By Eqn. (5), each factor in the product equals to either 1 or 0. Since some MAX conditions do not hold, by Lemma 1, the product does not equal to 1. That is, some factors of the product equal to 0, and so does the product.

□

Proof of Lemma 3

[Sufficiency] Assume a given tuple $(z_1,...,z_\omega,e)$ such that $e = MAX(z_1,...,z_\omega)$ holds. The sum of Eqn. (9) is over multiple terms, each of which is relative to a distinct tuple $(z_1,...,z_\omega,y_1,...,y_{\omega-2},e)$ of Ψ. By Lemma 1, exactly one term has the value 1, where

$$y_1 = MAX(z_1,z_2),\ ...,\ y_{\omega-2} = MAX(z_1,...,z_{\omega-1}).$$

By Lemma 2, all other terms have the value 0. Hence, $P_\Psi(e|z_1,...,z_\omega) = 1$.

[Necessity] Assume a given tuple $(z_1,...,z_\omega,e)$ such that $P_\Psi(e|z_1,...,z_\omega) = 1$. Each term in the sum is a product relative to a distinct tuple $(z_1,...,z_\omega,y_1,...,y_{\omega-2},e)$ of Ψ. By Lemma 2, a term has the value 0 unless all of the following hold:

$$y_1 = MAX(z_1,z_2),\ ...,\ y_{\omega-2} = MAX(z_1,...,z_{\omega-1}),$$

$$e = MAX(z_1,...,z_\omega).$$

Since $P_\Psi(e|z_1,...,z_\omega) = 1$, not all terms are 0. By Lemma 1, exactly one may have value 1, whose tuple satisfies all MAX conditions above. Hence, $e = MAX(z_1,...,z_\omega)$. □

Proof of Lemma 4

Each term in the sum is relative to a distinct tuple $(z_1,...,z_\omega,y_1,...,y_{\omega-2},e)$ of Ψ. By Lemma 2, a term has the value 0 unless all of the following hold:

$$y_1 = MAX(z_1,z_2),\ ...,\ y_{\omega-2} = MAX(z_1,...,z_{\omega-1}),$$

$$e = MAX(z_1,...,z_\omega).$$

As the last MAX condition fails, the result follows. □

Proof of Theorem 3

Let Φ be the SBDu of the dual gate model with MAX CPT $P_\Phi(e|z_1,...,z_\omega)$. By Eqn. (5), $P_\Phi(e|z_1,...,z_\omega)$ equals to 1 whenever $e = MAX(z_1,...,z_\omega)$. By Lemma 3, if $e = MAX(z_1,...,z_\omega)$, then $P_\Psi(e|z_1,...,z_\omega) = 1 = P_\Phi(e|z_1,...,z_\omega)$.

By Eqn. (5), $P_\Phi(e|z_1,...,z_\omega)$ equals to 0 whenever $e \neq MAX(z_1,...,z_\omega)$. By Lemma 4, if $e \neq MAX(z_1,...,z_\omega)$, then $P_\Psi(e|z_1,...,z_\omega) = 0 = P_\Phi(e|z_1,...,z_\omega)$.

Hence, the CPT of DSDu Ψ satisfies $P_\Psi(e|z_1,...,z_\omega) = P_\Phi(e|z_1,...,z_\omega)$. From Theorem 1, the result follows. □

Proof of Theorem 4

Let Φ be the SBDi of the direct gate model with PMIN CPT $P_\Phi(e|z_1,...,z_\omega)$ by Eqn. (8). We need to show $P_\Psi(e|z_1,...,z_\omega) = P_\Phi(e|z_1,...,z_\omega)$.

Theorem 4, relating SBDi and DSDi, is similar to Theorem 3, relating SBDu and DSDu. The main difference is that MAX CPTs (Eqn. (5)) in SBDu and DSDu are replaced by PMIN CPTs (Eqn. (8)) and PMIN$^+$ CPTs (Eqn. (10)) in SBDi and DSDi. The difference involves replacing MAX by MIN, and handling *aaci* values. Without *aaci* values, Theorem 3 would be trivially extended to direct gate model, due to symmetry between MAX and MIN. Hence, we focus on justifying that $P_\Psi(e|z_1,...,z_\omega)$ behaves according to Eqn. (8) when some z_i equals *aaci*.

First, we show that if each $z_i = aaci$ and $e = e^0$, then $P_\Psi(e|z_1,...,z_\omega) = 1$. Similarly to Lemma 1, consider the tuple $(z_1,...,z_\omega,y_1,...,y_{\omega-2},e)$, where each z_i and y_i equal *aaci*. It determines value of one term in the sum of Eqn. (9). By Eqn. (10), we have $P_\Psi(y_1 = aaci|z_1 = aaci, z_2 = aaci) = 1$, and $P_\Psi(y_i = aaci|y_{i-1} = aaci, z_{i+1} = aaci) = 1$. By Eqn. (8), we have $P_\Psi(e = e^0|y_{\omega-2} = aaci, z_\omega = aaci) = 1$. Hence, this term equals 1. Similarly to Lemma 2, by Eqns. (10) and (8), all other terms of the sum equal 0. Hence, similarly to Lemma 3, we conclude $P_\Psi(e|z_1,...,z_\omega) = 1$.

Second, we show without losing generality that, if $z_1,...,z_m = aaci$ $(1 \leq m < \omega)$, $z_{m+1},...,z_\omega \neq aaci$, and $e = MIN(z_{m+1},...,z_\omega)$, then $P_\Psi(e|z_1,...,z_\omega) = 1$. We consider $m = 1$ and $m > 1$ separately.

Suppose $m = 1$. Consider the tuple $(z_1,...,z_\omega,y_1,...,y_{\omega-2},e)$, where $z_1 = aaci$, $y_1 = z_2$, and $y_i = MIN(z_i,z_{i+1})$ $(i =$

$2,...,\omega-2$). Similarly to Lemma 1, by Eqn. (10), we have $P_\Psi(y_1|z_1,z_2)=1$ and $P_\Psi(y_i|y_{i-1},z_{i+1})=1$ $(i=2,...,\omega-2)$. By Eqn. (8), we have $P_\Psi(e|y_{\omega-2},z_\omega)=1$. Hence, $P_\Psi(e|z_1,...,z_\omega)=1$.

Next, suppose $m>1$. Consider $(z_1,...,z_\omega,y_1,...,y_{\omega-2},e)$, where $y_i=aaci$ $(i=1,...,m-1)$, $y_m=z_{m+1}$, and $y_i=MIN(z_i,z_{i+1})$ $(i=m+1,...,\omega-2)$. By Eqn. (10), $P_\Psi(y_1|z_1,z_2)=1$ and $P_\Psi(y_i|y_{i-1},z_{i+1})=1$ $(i=2,...,\omega-2)$. By Eqn. (8), $P_\Psi(e|y_{\omega-2},z_\omega)=1$. Hence, $P_\Psi(e|z_1,...,z_\omega)=1$.

Therefore, we have $P_\Psi(e|z_1,...,z_\omega)=P_\Phi(e|z_1,...,z_\omega)$. From Theorem 2, the result follows. □

Proof of Lemma 5

We prove each of the two cases. In the 1st case, we have $q=aaci$ and each $z_i=aaci$. Consider tuple $t=(z_1,...,z_\omega,y_1,...,y_{\omega-2},q)$, where each $y_i=aaci$. By PMAX$^+$ CPT in Eqn. (11), $P(y_1|z_1,z_2)=1$, $P(y_i|y_{i-1},z_{i+1})=1$ $(i=2,...,\omega-2)$, and $P(q|y_{\omega-2},z_\omega)=1$. Hence, $Q(t)=1$ by Eqn. (9). From Eqn. (11), for each tuple t' where not every y_i equals $aaci$, $Q(t')=0$. Hence, $P_\Psi(q|z_1,...,z_\omega)=1$ in the 1st case.

In the 2nd case, without losing generality, assume that $z_1,...,z_m\neq aaci$ $(1\leq m\leq\omega)$ and $z_{m+1},...,z_\omega=aaci$. Consider tuple $t=(z_1,...,z_\omega,y_1,...,y_{\omega-2},q)$, where $y_i=MAX(z_1,...,z_{i+1})$ $(i=1,...,m-1)$, $y_j=MAX(z_1,...,z_m)$ $(j=m,...,\omega-2)$, and $q=MAX(z_1,...,z_m)$. Applying Eqn. (11) to each factor of function $Q(t)$, where the $aaci$ argument has no effect, we have $Q(t)=1$. For every other tuple t' where some of the above conditions on t does not hold, we have $Q(t)=0$. Hence, $P_\Psi(q|z_1,...,z_\omega)=1$ in the 2nd case when $q=MAX(z_1,...,z_m)$.

By Eqn. (5), when $e=MAX(z_1,...,z_m)$, we have $P_\Phi(e|z_1,...,z_\omega)=1$. Hence, $P_\Psi(q|z_1,...,z_\omega)=P_\Phi(e|z_1,...,z_\omega)$ in the 2nd case. □

Proof of Theorem 5

Let Φ be the SBDu of the dual gate model. Assume that some W_i are active. By Lemma 5, $P_\Psi(q|z_1,...,z_\omega)$ is equivalent to the MAX CPT of SBDu Φ. From Theorem 1, it follows that $P_\Psi(q|z_1,...,z_\omega)$ equals $P(e|W_1,...,W_\omega)$ of the dual gate model.

Next, assume that all W_i are inactive. $P_\Psi(q|W_1,...,W_\omega)$ sums multiple products. Each product is relative to a tuple $(z_1,...,z_\omega,q)$. Consider the tuple t where each $z_i=aaci$ and $q=aaci$. By Eqn. (6), $P_\Psi(z_i=aaci|c_{i1},...,c_{i\theta_i})=1$ holds for each z_i since W_i is inactive. By Lemma 5, $P_\Psi(q=aaci|z_1,...,z_\omega)=1$. Hence, the product relative to t equals 1. For each t' where some $z_i\neq aaci$, by Eqn. (6), $P_\Psi(z_i|c_{i1},...,c_{i\theta_i})=0$. Hence, the product relative to t' equals 0. Therefore, $P_\Psi(q=aaci|W_1,...,W_\omega)=1$. □

Proof of Theorem 6

We prove by extending Theorem 5. Variables z_i, y_i, and q in DEDu and DEDi have the same domain. CPTs of z_i variables in both follow Eqn. (6). In a DEDu, y_i and q have PMAX$^+$ CPTs of Eqn. (11). In a DEDi, y_i and q have PMIN$^+$ CPTs of Eqn. (10). The two equations are symmetric.

Theorem 5 concerns dual gate models and is derived through Theorem 1. From the symmetry between dual gate models and direct gate models, and that between Theorem 1 and Theorem 2, the result follows from Theorem 5. □

Proof of Theorem 8

We show the equation from the left-hand side to the right-hand side. Denote $V=X\cup Y$, where $X\cap Y=\emptyset$, X consists of root and child variables whose CPTs are in TC, and Y consists of effect variables of NAT models in NM. By the chain rule of BN, we have

$$P_\Omega(V,W)=\big(\Pi_{x\in X}P_\Omega(x|\pi(x))\big)\big(\Pi_{v\in Y\cup W}P_\Omega(v|\pi(v))\big).$$

Denote $Y=\{e_1,...,e_k\}$, where $k=|NM|$ counts NAT models in Γ. Auxiliary and quasi-effect variables over BN segments of different NAT models are disjoint. Hence, we index subsets of variables in W as $W_1,...,W_k$, where $W_i\cap W_j=\emptyset$ $(i\neq j)$ and $\cup_i W_i=W$, such that W_i is the set of auxiliary and quasi-effect variables in the BN segment for NAT model over e_i.

It follows that

$$P_\Omega(V,W)=\big(\Pi_{x\in X}P_\Omega(x|\pi(x))\big)\Big(\Pi_{i=1}^k\big(\Pi_{v\in\{e_i\}\cup W_i}P_\Omega(v|\pi(v))\big)\Big).$$

The left-hand side of equation in the theorem becomes

$$\begin{aligned}&\sum_{w\in W}P_\Omega(V,W)\\&=\big(\Pi_{x\in X}P_\Omega(x|\pi(x))\big)\sum_{w\in W}\Big(\Pi_{i=1}^k\big(\Pi_{v\in\{e_i\}\cup W_i}P_\Omega(v|\pi(v))\big)\Big)\\&=\big(\Pi_{x\in X}P_\Omega(x|\pi(x))\big)\Pi_{i=1}^k\Big(\sum_{w\in W_i}\big(\Pi_{v\in\{e_i\}\cup W_i}P_\Omega(v|\pi(v))\big)\Big).\end{aligned}$$

The first equality holds since $X\cap W=\emptyset$. The second equality holds since $W_i\cap W_j=\emptyset$ $(i\neq j)$. In the last expression, each summation is the marginalized product of CPTs in the BN segment for NAT model over e_i. By Theorem 7, the summation is equal to $P_\Gamma(e_i|\pi(e_i))$: the CPT of the NAT model over e_i defined by Γ.

Since TC in Γ and TC in Ω are identical, we have

$$\begin{aligned}\sum_{w\in W}P_\Omega(V,W)&=\big(\Pi_{x\in X}P_\Gamma(x|\pi(x))\big)\Pi_{i=1}^k P_\Gamma(e_i|\pi(e_i))\\&=\Pi_{v\in V}P_\Gamma(v|\pi(v))=P_\Gamma(V).\end{aligned}$$

□