Regular n-ary Queries in Trees and Variable Independence

Emmanuel Filiot Sophie Tison

Laboratoire d'Informatique Fondamentale de Lille (LIFL)
INRIA Lille Nord-Europe, Mostrare Project

IFIP TCS, 2008

Motivations

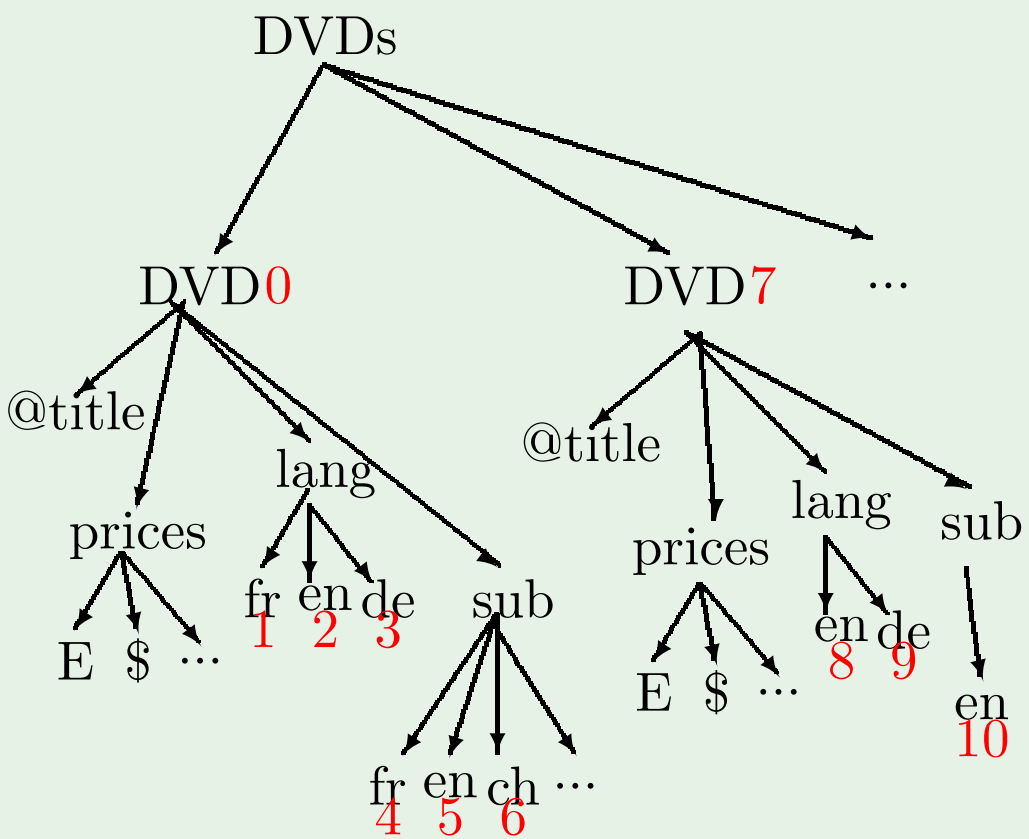
- n-ary queries $\phi(x_1, \ldots, x_n)$ in trees t: select n-tuples of nodes
- fundamental to XML processing tasks
- the set of answers can grow exponentially in $|t|$ ($O(|t|^n)$ in the worst case)
- the answers share common information
- a compact representation is needed

Aggregated Answers (Meuss, Schulz, Bry, ICDT'01)

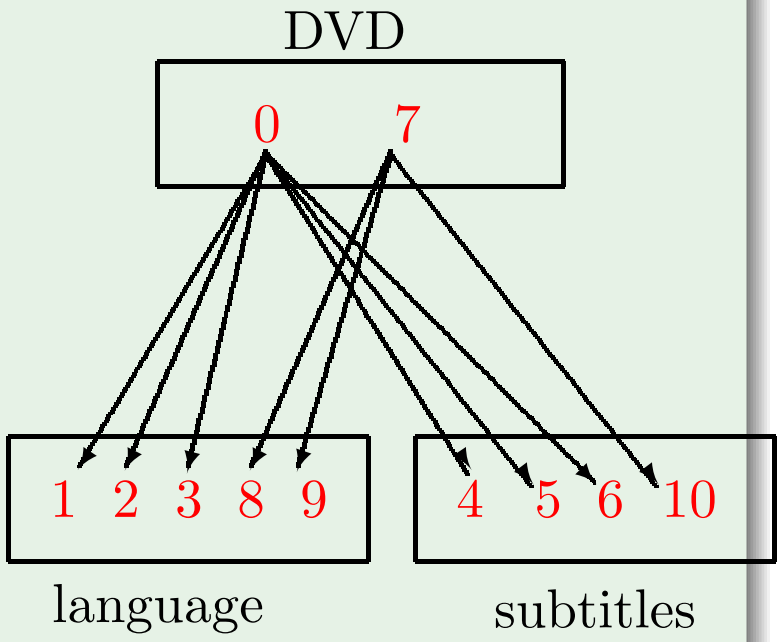
- join-free term over $\{\times, \vee\}$ and constants $[x \mapsto n]$;

Example (aggregated answers)

XML document



Aggregated Answer



Query: (DVD, Lang, Subtitles)

Advantages of this representation

- efficient model-checking: $\tau \in Ans(\phi, t)$?
- efficient enumeration: $\tau_1, \ldots, \tau_i, \ldots$
- advanced query answering:
  - answer searching, statistics
  - answer browsing
  - cascade style query-answering
- view computing

Objective

How compact are aggregated answers? How to measure compactness?

Objective

How compact are aggregated answers? How to measure compactness?

- compactness is related to variable independence
- variable independence in the context of infinite constraint databases (Chomicki et. al., PODS'96) (Libkin, TOCL'03)
- only between blocks of variables, and fixed database
- we propose a more general notion: dependency forests
- MSO in ordered finite binary trees

Objective

How compact are aggregated answers? How to measure compactness?

- compactness is related to variable independence
- variable independence in the context of infinite constraint databases (Chomicki et. al., PODS'96) (Libkin, TOCL'03)
- only between blocks of variables, and fixed database
- we propose a more general notion: dependency forests
- MSO in ordered finite binary trees

Question

Given a dependency forest, and a query, are the variables independent w.r.t. this forest?

Outline

1. **variable independence w.r.t. a partition**
2. variable independence w.r.t. a dependency forest

Trees and MSO

- finite ordered **binary** trees t are viewed as structures over the signature $S_1(x,y), S_2(x,y), \mathrm{lab}_a(x)$, $a \in \Sigma$, with domain $nodes(t)$
- first-order variables $x, y, z, \ldots$ denote **nodes**
- second-order variables $X, Y, Z, \ldots$ denote **set of nodes**
- existential quantifiers $\exists x$, $\exists X$, Boolean operators $\neg, \vee, \wedge$, and membership $x \in X$;
- an MSO formula $\phi(x_1, \ldots, x_n)$ defines an n-ary query:

$$Ans(\phi, t) \;=\; \{(u_1, \ldots, u_n) \in nodes(t)^n \mid t \models \phi(u_1, \ldots, u_n)\}$$

Trees and MSO

- finite ordered **binary** trees t are viewed as structures over the signature $S_1(x,y), S_2(x,y), \mathrm{lab}_a(x)$, $a \in \Sigma$, with domain $nodes(t)$
- first-order variables $x, y, z, \ldots$ denote **nodes**
- second-order variables $X, Y, Z, \ldots$ denote **set of nodes**
- existential quantifiers $\exists x$, $\exists X$, Boolean operators $\neg, \vee, \wedge$, and membership $x \in X$;
- an MSO formula $\phi(x_1, \ldots, x_n)$ defines an n-ary query:

$$Ans(\phi, t) \;=\; \{(u_1, \ldots, u_n) \in nodes(t)^n \mid t \models \phi(u_1, \ldots, u_n)\}$$

Example

Select all the DVDs:

$$\phi(x) \;=\; lab_{DVD}(x)$$

Trees and MSO

- finite ordered **binary** trees t are viewed as structures over the signature $S_1(x,y), S_2(x,y), \mathrm{lab}_a(x)$, $a \in \Sigma$, with domain $nodes(t)$
- first-order variables $x, y, z, \ldots$ denote **nodes**
- second-order variables $X, Y, Z, \ldots$ denote **set of nodes**
- existential quantifiers $\exists x$, $\exists X$, Boolean operators $\neg, \vee, \wedge$, and membership $x \in X$;
- an MSO formula $\phi(x_1, \ldots, x_n)$ defines an n-ary query:

$$Ans(\phi, t) \;=\; \{(u_1, \ldots, u_n) \in nodes(t)^n \mid t \models \phi(u_1, \ldots, u_n)\}$$

Example

Select all the triples $(dvd, lang, sub)$:

$$\begin{aligned} \phi(x,y,z) \;=\; & lab_{DVD}(x) \wedge lab_{lang}(y) \wedge lab_{sub}(z) \wedge \\ & desc(x,y) \wedge desc(x,z) \end{aligned}$$

Variable Independence w.r.t. a Partition

- **input**: a formula $\phi(x_1, \ldots, x_n)$, a partition $P = \{B_1, \ldots, B_k\}$ of $\{x_1, \ldots, x_n\}$
- **output**: is ϕ equivalent to a formula of the form:

$$\bigvee_{i=1}^{N} \phi_i^1(B_1) \wedge \cdots \wedge \phi_i^k(B_k)$$

where $freevar(\phi_i^j) = B_j$. We say that ϕ **conforms** to P.

Theorem

Variable independence w.r.t. a partition is decidable, and a decomposition is computable.

- For all $i \in \{1, \ldots, k\}$, consider:

$$\begin{aligned} Swap_i(\overline{x}, \overline{y}) \quad = \quad & \forall B_1 \ldots \forall B_{i-1} \forall B_{i+1} \ldots \forall B_n \\ & \phi(B_1, \ldots, B_{i-1}, \ \overline{x} \ , B_{i+1}, \ldots, B_n) \leftrightarrow \\ & \phi(B_1, \ldots, B_{i-1}, \ \overline{y} \ , B_{i+1}, \ldots, x_{B_n}) \end{aligned}$$

- For all $i \in \{1, \ldots, k\}$, consider:

$$\begin{aligned} Swap_i(\overline{x}, \overline{y}) \quad = \quad & \forall B_1 \ldots \forall B_{i-1} \forall B_{i+1} \ldots \forall B_n \\ & \phi(B_1, \ldots, B_{i-1}, \ \overline{x} \ , B_{i+1}, \ldots, B_n) \leftrightarrow \\ & \phi(B_1, \ldots, B_{i-1}, \ \overline{y} \ , B_{i+1}, \ldots, x_{B_n}) \end{aligned}$$

- intuition: if $t \models Swap_i(\overline{u}, \overline{v})$ then $\overline{u}$ and $\overline{v}$ are not distinguished by ϕ.

- For all $i \in \{1, \ldots, k\}$, consider:

$$\begin{aligned} Swap_i(\overline{x}, \overline{y}) \quad = \quad & \forall B_1 \ldots \forall B_{i-1} \forall B_{i+1} \ldots \forall B_n \\ & \phi(B_1, \ldots, B_{i-1}, \ \overline{x} \ , B_{i+1}, \ldots, B_n) \leftrightarrow \\ & \phi(B_1, \ldots, B_{i-1}, \ \overline{y} \ , B_{i+1}, \ldots, x_{B_n}) \end{aligned}$$

- intuition: if $t \models Swap_i(\overline{u}, \overline{v})$ then $\overline{u}$ and $\overline{v}$ are not distinguished by ϕ.
- Every formula $Swap_i(\overline{x}, \overline{y})$ defines on a tree t an equivalence relation $Ans(Swap_i, t)$ between tuples of size $|B_i|$.

Variable Independence Reduces to Query Boundedness

Theorem

ϕ conforms to $\{B_1, \ldots, B_k\}$ iff for all $i = 1, \ldots, k$, $Swap_i$ is of bounded index, i.e.:

$$\exists b_i \geq 0,\ \forall t,\ \left|nodes(t)^{|B_i|}/_{Ans(Swap_i,t)}\right| \leq b_i$$

The decomposition has the following form:

$$\phi(x_1, \ldots, x_n) \ \leftrightarrow\ \bigvee_j \phi_j \wedge cl_1^j(B_1) \wedge \cdots \wedge cl_k^j(B_k)$$

On Deciding Boundedness

Theorem

Given a formula $\phi(x)$, we can decide if there is a bound $b \geq 0$ such that:

$$\forall t,\ |Ans(\phi, t)| \leq b$$

Moreover, some bound b is computable.

1. translate $\phi(x)$ into a canonical $\{0,1\}$-labeling transducer T_ϕ
2. decide whether the number of images of any tree by T_ϕ is bounded by some constant [Seidl, habilitation thesis].

Extension to n-ary queries

$\phi(x_1, \ldots, x_n)$ is bounded iff for each i, the following is bounded:
$proj_i(x) = \exists x_1 \ldots \exists x_{i-1} \exists x_{i+1} \ldots \exists x_n \phi(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$

Bounded Index Property

Corollary

Bounded index property is decidable for every $Swap_i(\overline{x}, \overline{y})$*, moreover, an index is computable.*

1. define a total order $\overline{x} \leq^n \overline{y}$ on n-tuples of nodes (by an MSO formula);
2. decide boundedness of $Min(\overline{x})$, which selects the minimal representatives of the relation defined by $Swap_i(\overline{x}, \overline{y})$.

Orthographic Dimension

Definition (Grumbach, Rigaux, Segoufin, ICDT'99)

Let $\phi(x_1, \ldots, x_n)$ be a formula, and $\mathcal{P}$ the set of partitions P such that ϕ conforms to P. The orthographic dimension $d(\phi)$ is defined by:

$$d(\phi) \;=\; min_{\{B_1,\ldots,B_k\}\in\mathcal{P}}\; max_i\; |B_i|$$

- how to compute it? try every partition of $\{x_1, \ldots, x_n\}$.
- **improvement**: consider only 2-partitions, thanks to the following theorem, adapted from (Cosmadakis, Kuper, Libkin, 01):

Orthographic Dimension

Definition (Grumbach, Rigaux, Segoufin, ICDT'99)

Let $\phi(x_1, \ldots, x_n)$ be a formula, and $\mathcal{P}$ the set of partitions P such that ϕ conforms to P. The orthographic dimension $d(\phi)$ is defined by:

$$d(\phi) \;=\; min_{\{B_1,\ldots,B_k\}\in\mathcal{P}}\; max_i\; |B_i|$$

- how to compute it? try every partition of $\{x_1, \ldots, x_n\}$.
- **improvement**: consider only 2-partitions, thanks to the following theorem, adapted from (Cosmadakis, Kuper, Libkin, 01):

Theorem

If ϕ conforms to P_1 and P_2, then ϕ conforms to $P_1 \sqcap P_2$.

Relation to answer set representation

We can compute a formula equivalent to $\phi(x_1, \ldots, x_n)$ which corresponds to the orthographic dimension, i.e. a formula

$$\bigvee_{i=1}^{N} \phi_i^1(B_1) \wedge \cdots \wedge \phi_i^k(B_k) \qquad \text{where } d(\phi) = max_i \ |B_i|$$

Relation to answer set representation

We can compute a formula equivalent to $\phi(x_1, \ldots, x_n)$ which corresponds to the orthographic dimension, i.e. a formula

$$\bigvee_{i=1}^{N} \phi_i^1(B_1) \wedge \cdots \wedge \phi_i^k(B_k) \qquad \text{where } d(\phi) = max_i\ |B_i|$$

The aggregated answer on a tree t is of size $O(f(|\phi|).|t|^{d(\phi)})$.

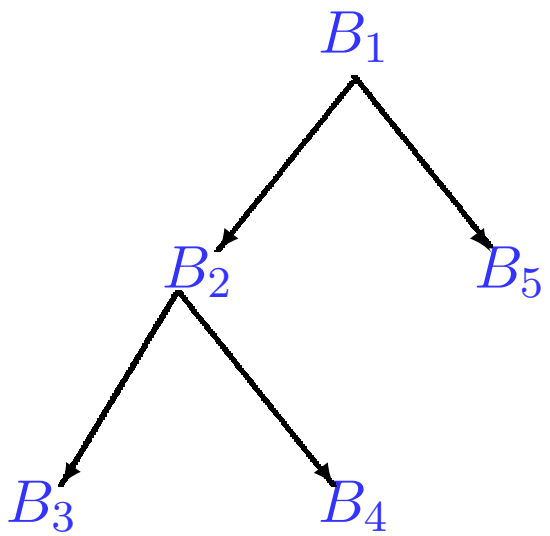
- compute the answer sets A_i^j of each ϕ_i^j
- represent the answer set by a union (whose size only depends on ϕ, which is fixed) of cartesian products $A_i^1 \times \cdots \times A_i^k$.

Outline

1. variable independence w.r.t. a partition
2. **variable independence w.r.t. a dependency forest**

Dependency Forest F

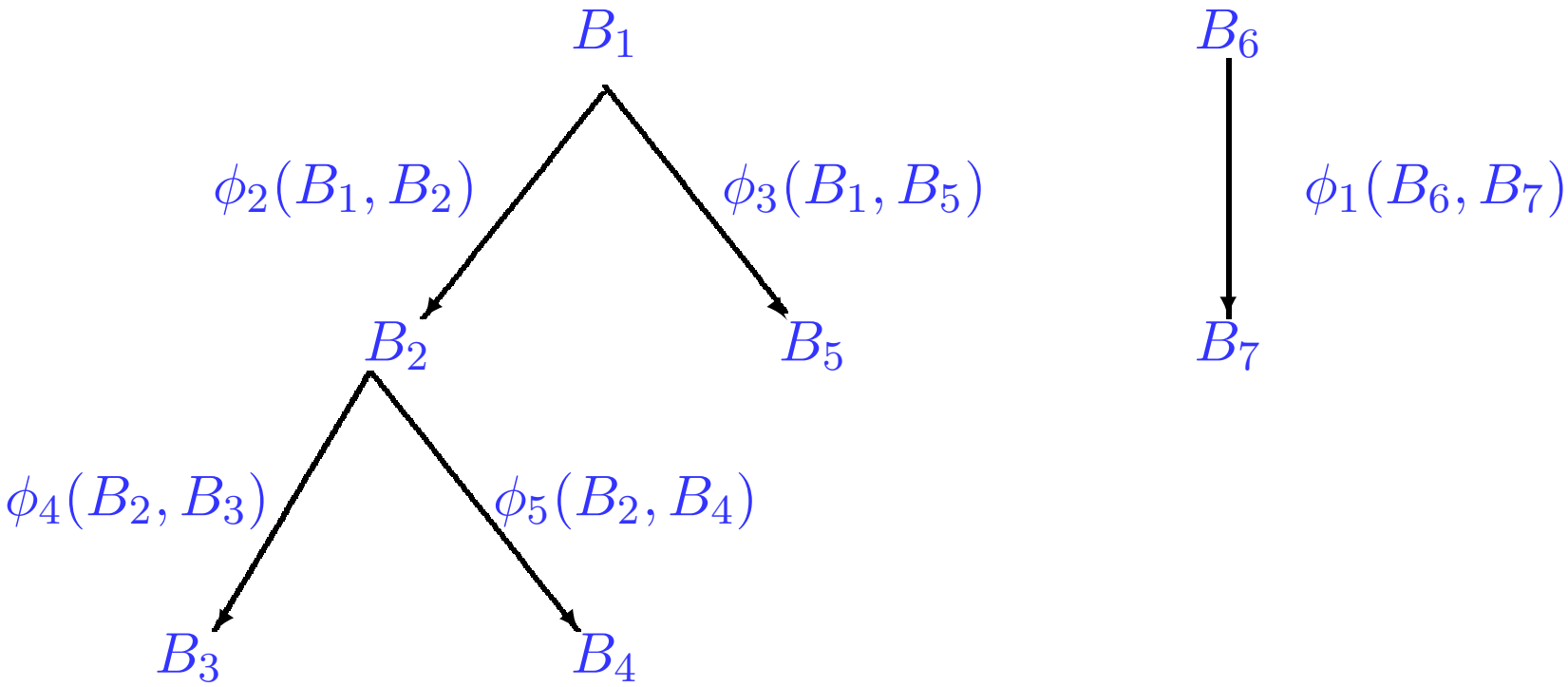
- over a set $V = \{x_1, \ldots, x_n\}$ of variables



where $\{B_1, \ldots, B_7\}$ partitions V

Conformance to a Dependency Forest F

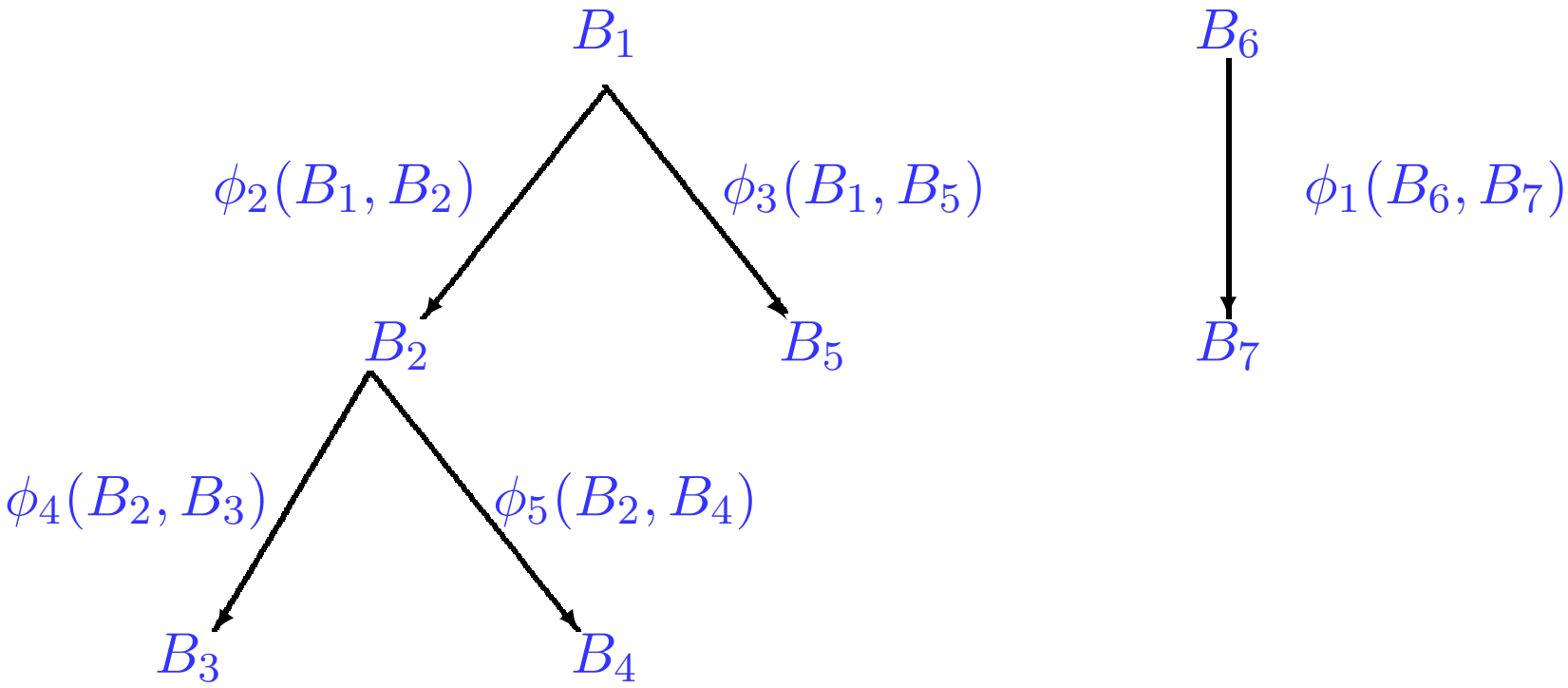
- $\{B_1, \ldots, B_7\}$ partitions $V = \{x_1, \ldots, x_n\}$



- μ : edges(F) $\rightarrow$ MSO formulas
- $\mu(F) =$
  $\phi_1(B_6, B_7) \wedge \phi_2(B_1, B_3) \wedge \phi_3(B_1, B_5) \wedge \phi_4(B_2, B_3) \wedge \phi_5(B_2, B_5)$

Conformance to a Dependency Forest F

- $\{B_1, \ldots, B_7\}$ partitions $V = \{x_1, \ldots, x_n\}$



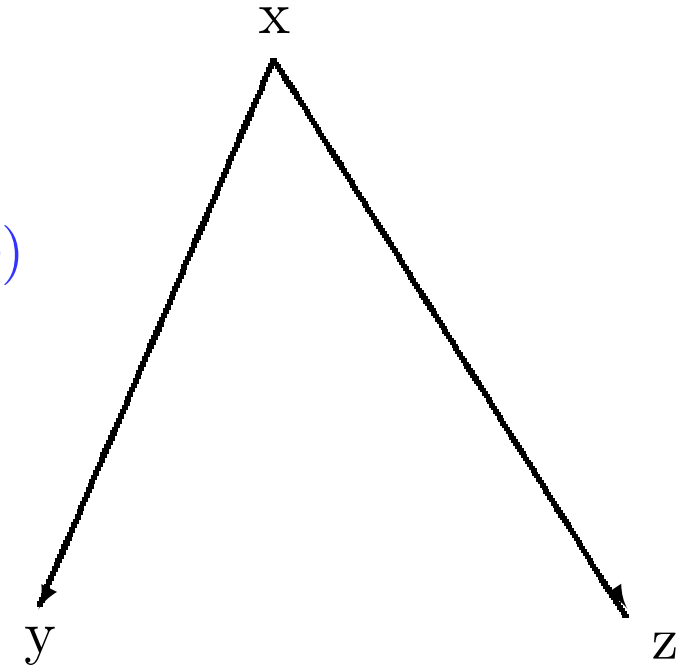
Definition

A formula $\phi(x_1, \ldots, x_n)$ conforms to F if there is a finite sequence $\mu_1, \ldots, \mu_k$ such that $\phi \leftrightarrow \bigvee_i \mu_i(F)$.

Example

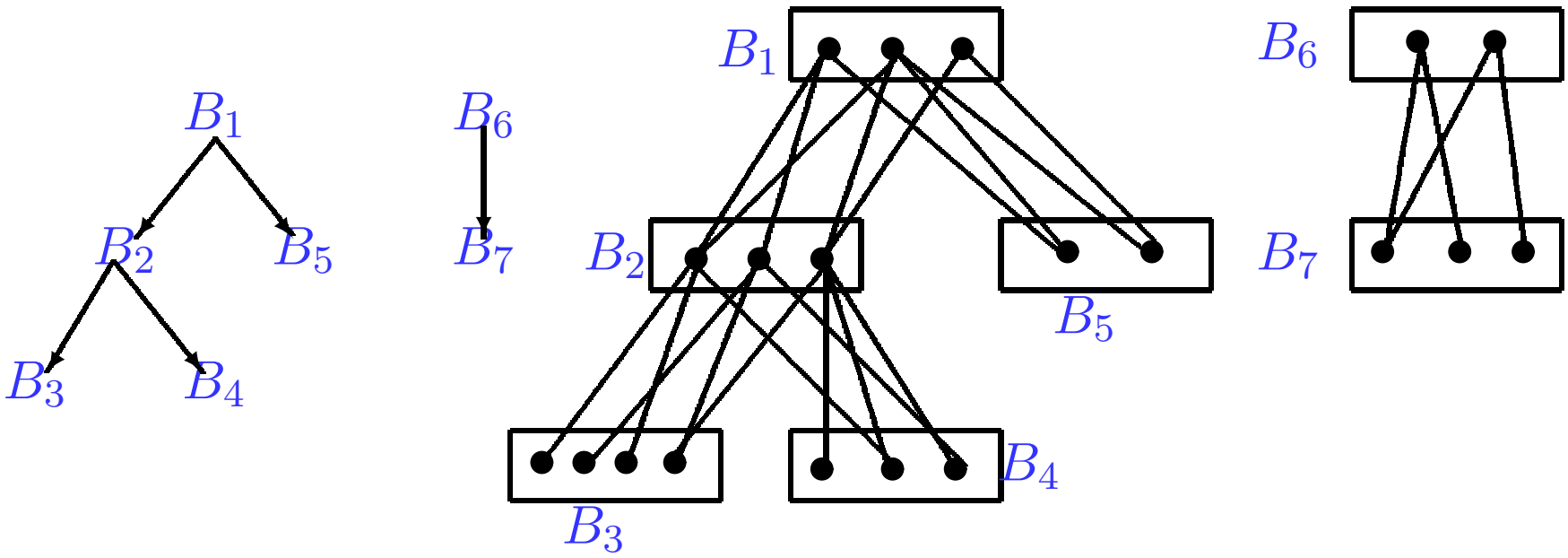
The query $\phi(x, y, z) \ : \ t \mapsto (DVD, Lang, Sub)$

conforms to



Relation to Answer Set Representation

The set of answers can be represented by an aggregated answer of size $O(f(|\phi|).|t|^{2b})$, where $b = max_i\ |\overline{y}_i|$.



Main Result

Theorem

It is decidable whether a formula $\phi(x_1, \ldots, x_n)$ conforms to a dependency forest F.

1. decide it for forests of the form $B_1(B_2, B_3)$ $\rightarrow$ signature $\Sigma \times \{0,1\}^{B_1}$

Main Result

Theorem

It is decidable whether a formula $\phi(x_1, \ldots, x_n)$ conforms to a dependency forest F.

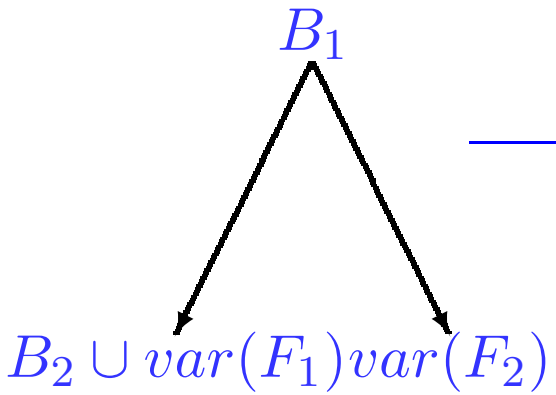
1. decide it for forests of the form $B_1(B_2, B_3) \rightarrow$ signature $\Sigma \times \{0,1\}^{B_1}$
2. inductively: if $F = \{T_1, \ldots, T_k\}$
   - decompose ϕ w.r.t. the partition $\{var(T_1), \ldots, var(T_k)\}$
   - get a disjunction of the form $\bigvee_i \alpha_i^1(var(T_1)) \wedge \cdots \wedge \alpha_i^k(var(T_k))$
   - decompose each α_i^j w.r.t. T_j

Main Result

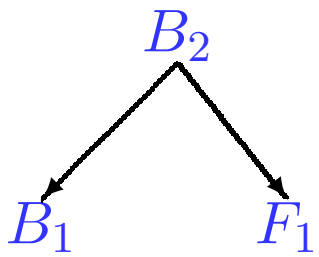
Theorem

It is decidable whether a formula $\phi(x_1, \ldots, x_n)$ conforms to a dependency forest F.

1. decide it for forests of the form $B_1(B_2, B_3)$ $\rightarrow$ signature $\Sigma \times \{0,1\}^{B_1}$
2. inductively: if $F = B_1(B_2(F_1), F_2)$:



$$\longrightarrow \bigvee_i \phi_1^i(B_1,\ B_2 \cup var(F_1)) \ \wedge \ \phi_2^i(B_1, var(F_2))$$



Are tuples of variables really needed in F?

Proposition

*There is an MSO formula ϕ which **does not conform** to any dependency forest whose nodes are labeled by **single variables**.*

- take $\phi(x, y, z)$ defined by $lca(x, y) = lca(x, z)$
- find counter-examples for every forest F over $\{x, y, z\}$

Perspectives

- **extend the structures beyond trees**
  - need an MSO-definable total order
  - decidability of boundedness
  - $\Longrightarrow$ variable independence is decidable for unranked trees

Perspectives

- **extend the structures beyond trees**
  - need an MSO-definable total order
  - decidability of boundedness
  - $\Longrightarrow$ variable independence is decidable for unranked trees
- **tree pattern queries**
  - n-ary tree patterns with $desc$, $child$, $next-sibling$, $label$ tests
  - fragments for which the aggregated answers have size $O(poly(|\phi|).|t|)$