Aggregation in Probabilistic Databases via Knowledge Compilation

Robert Fink, Larisa Han, Dan Olteanu
University of Oxford

VLDB 2012, Istanbul
Outline

Motivation

Algebraic Foundations

Representation System

Query Evaluation
Who is responsible for a larger capacity of biogas plants, Democrats or Republicans?
Who is responsible for a larger capacity of biogas plants, Democrats or Republicans?
Who is responsible for a larger capacity of biogas plants, Democrats or Republicans?
More biomass plant capacity, Democrats or Republicans?

How to come up with an answer?

- Option 1: Use Wikipedia, search for lists of Governors and their terms. Search for list of biomass plants, find out when and where they were build, match up with Governors of US states. Group by political parties of Governors, sum capacity of plants. (Phew.)
More biomass plant capacity, Democrats or Republicans?

How to come up with an answer?

- Option 1: Use Wikipedia, search for lists of Governors and their terms. Search for list of biomass plants, find out when and where they were built, match up with Governors of US states. Group by political parties of Governors, sum capacity of plants. (Phew.)

- Option 2: Find tables on Governors and biomass plants on the Web and write a query like

  ```sql
  compute sum(Plant.capacity) from Governor, Plant where
  - Plant.date matches Governor.term
  - Plant.location matches Governor.state
  group by Governor.party
  ```
## Biomass Plants in the US

### Add a new Energy Generation Facility

![Map of Biomass Plants in the US](https://via.placeholder.com/150)

### List of Biomass Facilities
- AES Mendota Biomass Facility
- APS Biomass I Biomass Facility
- Aberdeen Biomass Facility
- Acme Landfill Biomass Facility
- Adrian Energy Associates LLC Biomass Facility
- Agrilelectric Power Partners Ltd Biomass Facility
- Al Turi Biomass Facility
- Alabama Pine Pulp Biomass Facility
- Albany Landfill Gas Utilization Project Biomass Facility
- Alexandria Biomass Facility
- Altamont Gas Recovery Biomass Facility
- American Canyon Power Plant Biomass Facility
- American Ref-Fuel of Delaware Valley

### Table: Alabama Pine Pulp Biomass Facility

<table>
<thead>
<tr>
<th>Name</th>
<th>Alabama Pine Pulp Biomass Facility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facility</td>
<td>Alabama Pine Pulp</td>
</tr>
<tr>
<td>Sector</td>
<td>Biomass</td>
</tr>
<tr>
<td>Location</td>
<td>Monroe County, Alabama</td>
</tr>
<tr>
<td>Coordinates</td>
<td>31.5119068°, -87.460397°</td>
</tr>
<tr>
<td>Generating Capacity (MW)</td>
<td>32.085</td>
</tr>
<tr>
<td>Commercial Online Date</td>
<td>1991</td>
</tr>
<tr>
<td>Heat Rate (BTU/LW)</td>
<td>1580.0 C</td>
</tr>
</tbody>
</table>

Map data ©2011 Europa Technologies, INEGI.
## Past Governors Bios

### Status

- All
- In Office
- Out of Office

### State

- All States
- Alabama
- Alaska

<table>
<thead>
<tr>
<th>Governor's Name</th>
<th>State</th>
<th>Time in Office</th>
<th>Party</th>
</tr>
</thead>
</table>
### Deterministic case

<table>
<thead>
<tr>
<th>G.Name</th>
<th>G.Party</th>
<th>G.State</th>
<th>P.Location</th>
<th>P.capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>Dem</td>
<td>CA</td>
<td>CA</td>
<td>17</td>
</tr>
<tr>
<td>G2</td>
<td>Dem</td>
<td>FL</td>
<td>FL</td>
<td>5</td>
</tr>
<tr>
<td>G3</td>
<td>Dem</td>
<td>NY</td>
<td>NY</td>
<td>9</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G4</td>
<td>Rep</td>
<td>NY</td>
<td>NY</td>
<td>8</td>
</tr>
<tr>
<td>G5</td>
<td>Rep</td>
<td>CA</td>
<td>CA</td>
<td>14</td>
</tr>
<tr>
<td>G6</td>
<td>Rep</td>
<td>CA</td>
<td>CA</td>
<td>2</td>
</tr>
</tbody>
</table>

Problem to solve: $17 + 5 + 9 > 8 + 14 + 2$?
Uncertain case

<table>
<thead>
<tr>
<th>G.Name</th>
<th>G.Party</th>
<th>G.State</th>
<th>P.Location</th>
<th>P.capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Dem</td>
<td>CA</td>
<td>SF, CA</td>
<td>17</td>
</tr>
<tr>
<td>P2</td>
<td>Dem</td>
<td>FL</td>
<td>Florida</td>
<td>5</td>
</tr>
<tr>
<td>P3</td>
<td>Dem</td>
<td>NY</td>
<td>NY</td>
<td>9</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P4</td>
<td>Rep</td>
<td>NY</td>
<td>NY</td>
<td>8</td>
</tr>
<tr>
<td>P5</td>
<td>Rep</td>
<td>CA</td>
<td>LA, CA</td>
<td>14</td>
</tr>
<tr>
<td>P6</td>
<td>Rep</td>
<td>CA</td>
<td>Berkeley</td>
<td>2</td>
</tr>
<tr>
<td>G.Name</td>
<td>G.Party</td>
<td>G.State</td>
<td>P.Location</td>
<td>P.capacity</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>P1</td>
<td>Dem</td>
<td>CA</td>
<td>SF, CA</td>
<td>17</td>
</tr>
<tr>
<td>P2</td>
<td>Dem</td>
<td>FL</td>
<td>Florida</td>
<td>5</td>
</tr>
<tr>
<td>P3</td>
<td>Dem</td>
<td>NY</td>
<td>NY</td>
<td>9</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P4</td>
<td>Rep</td>
<td>NY</td>
<td>NY</td>
<td>8</td>
</tr>
<tr>
<td>P5</td>
<td>Rep</td>
<td>CA</td>
<td>LA, CA</td>
<td>14</td>
</tr>
<tr>
<td>P6</td>
<td>Rep</td>
<td>CA</td>
<td>Berkeley</td>
<td>2</td>
</tr>
</tbody>
</table>
**Uncertain case**

<table>
<thead>
<tr>
<th>G.Name</th>
<th>G.Party</th>
<th>G.State</th>
<th>P.Location</th>
<th>P.capacity</th>
<th>Φ</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Dem</td>
<td>CA</td>
<td>SF, CA</td>
<td>17</td>
<td>$x_1$ (p=0.9)</td>
</tr>
<tr>
<td>P2</td>
<td>Dem</td>
<td>FL</td>
<td>Florida</td>
<td>5</td>
<td>$x_2$ (p=0.5)</td>
</tr>
<tr>
<td>P3</td>
<td>Dem</td>
<td>NY</td>
<td>NY</td>
<td>9</td>
<td>$x_3$ (p=1.0)</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P4</td>
<td>Rep</td>
<td>NY</td>
<td>NY</td>
<td>8</td>
<td>$y_1$ (p=1.0)</td>
</tr>
<tr>
<td>P5</td>
<td>Rep</td>
<td>CA</td>
<td>LA, CA</td>
<td>14</td>
<td>$y_2$ (p=0.8)</td>
</tr>
<tr>
<td>P6</td>
<td>Rep</td>
<td>CA</td>
<td>Berkeley</td>
<td>2</td>
<td>$y_3$ (p=0.2)</td>
</tr>
</tbody>
</table>
## Uncertain case

<table>
<thead>
<tr>
<th>G.Name</th>
<th>G.Party</th>
<th>G.State</th>
<th>P.Location</th>
<th>P.capacity</th>
<th>$\Phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Dem</td>
<td>CA</td>
<td>SF, CA</td>
<td>17</td>
<td>$x_1$ (p=0.9)</td>
</tr>
<tr>
<td>P2</td>
<td>Dem</td>
<td>FL</td>
<td>Florida</td>
<td>5</td>
<td>$x_2$ (p=0.5)</td>
</tr>
<tr>
<td>P3</td>
<td>Dem</td>
<td>NY</td>
<td>NY</td>
<td>9</td>
<td>$x_3$ (p=1.0)</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P4</td>
<td>Rep</td>
<td>NY</td>
<td>NY</td>
<td>8</td>
<td>$y_1$ (p=1.0)</td>
</tr>
<tr>
<td>P5</td>
<td>Rep</td>
<td>CA</td>
<td>LA, CA</td>
<td>14</td>
<td>$y_2$ (p=0.8)</td>
</tr>
<tr>
<td>P6</td>
<td>Rep</td>
<td>CA</td>
<td>Berkeley</td>
<td>2</td>
<td>$y_3$ (p=0.2)</td>
</tr>
</tbody>
</table>

**Problem to solve:**

$$x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 > y_1 \otimes 8 + y_2 \otimes 14 + y_3 \otimes 2$$
Algebraic Expressions give rise to Random Variables

Democratic Biogas Capacity   >   Republican Biogas Capacity

\[ 17 + 5 + 9 > 8 + 14 + 2 \]
Democratic Biogas Capacity $>$ Republican Biogas Capacity

\[
\Phi = [x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 > y_1 \otimes 8 + y_2 \otimes 14 + y_3 \otimes 2]
\]

- $x_1, x_2, x_3, y_1, y_2, y_3$ are Boolean random variables
Algebraic Expressions give rise to Random Variables

Democratic Biogas Capacity > Republican Biogas Capacity

\[ \Phi = [x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 > y_1 \otimes 8 + y_2 \otimes 14 + y_3 \otimes 2] \]

- \( x_1, x_2, x_3, y_1, y_2, y_3 \) are Boolean random variables
- Then the sum expression \( \alpha = x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 \) is an \( \mathbb{N} \)-valued random variable
Algebraic Expressions give rise to Random Variables

Democratic Biogas Capacity > Republican Biogas Capacity

\[ \Phi = [x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 > y_1 \otimes 8 + y_2 \otimes 14 + y_3 \otimes 2] \]

- \(x_1, x_2, x_3, y_1, y_2, y_3\) are Boolean random variables
- Then the sum expression \(\alpha = x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9\) is an \(\mathbb{N}\)-valued random variable
- Hence \(\Phi\) is a \(\mathbb{B}\)-valued random variable
Algebraic Expressions give rise to Random Variables

Democratic Biogas Capacity  >  Republican Biogas Capacity

\[ \Phi = [x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 > y_1 \otimes 8 + y_2 \otimes 14 + y_3 \otimes 2] \]

- \( x_1, x_2, x_3, y_1, y_2, y_3 \) are Boolean random variables
- Then the sum expression \( \alpha = x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 \) is an \( \mathbb{N} \)-valued random variable
- Hence \( \Phi \) is a \( \mathbb{B} \)-valued random variable
- \( P_\Phi[\top] \) is the probability that a random choice of possible values for the variables \( x_1, x_2, x_3, y_1, y_2, y_3 \) satisfies the inequality
Algebraic Expressions give rise to Random Variables

Democratic Biogas Capacity > Republican Biogas Capacity

\[ \Phi = [x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 > y_1 \otimes 8 + y_2 \otimes 14 + y_3 \otimes 2] \]

- \( x_1, x_2, x_3, y_1, y_2, y_3 \) are Boolean random variables
- Then the sum expression \( \alpha = x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 \) is an \( \mathbb{N} \)-valued random variable
- Hence \( \Phi \) is a \( \mathbb{B} \)-valued random variable
- \( P_\Phi[\top] \) is the probability that a random choice of possible values for the variables \( x_1, x_2, x_3, y_1, y_2, y_3 \) satisfies the inequality
- In previous example, \( P_\Phi[\top] \) is the probability that democrats were responsible for a higher capacity of biogas plants
Outline

Motivation

Algebraic Foundations

Representation System

Query Evaluation
Monoids, Semirings, Semimodule

What do we mean by $+$ in $\Phi_1 \otimes 17 + \Phi_2 \otimes 5$?
Well, it depends . . .
Monoids, Semirings, Semimodule

What do we mean by \(+\) in \(\Phi_1 \otimes 17 + \Phi_2 \otimes 5\)?
Well, it depends . . .

Aggregation modelled by commutative monoids

- Carrier \(M\), e.g. \(\mathbb{N}\) or \(\mathbb{R}\)
- Binary operation \(M \times M \rightarrow M\)
- Neutral element \(0 \in M\)
- Examples for aggregation monoids:
  - \(\text{SUM} (\mathbb{N}, +, 0)\), \(\text{MIN} (\mathbb{N}, \min, \infty)\), \(\text{MAX} (\mathbb{N}, \max, -\infty)\),
  - \(\text{PROD}\), \(\text{COUNT}\) (special case of \(\text{SUM}\))
Monoids, **Semirings**, Semimodule

What are $\Phi_1, \Phi_2$ in $\Phi_1 \otimes 17 + \Phi_2 \otimes 5$?
Monoids, **Semirings**, Semimodule

What are $\Phi_1$, $\Phi_2$ in $\Phi_1 \otimes 17 + \Phi_2 \otimes 5$?

Consider Query:

$\text{AGG}_B\left[ (R \cup S) \otimes_A T \right]$
Monoids, **Semirings**, Semimodule

What are $\Phi_1, \Phi_2$ in $\Phi_1 \otimes 17 + \Phi_2 \otimes 5$?

- Consider Query:

$$\text{AGG}_B \left[ (R \cup S) \Join_A T \right]$$

<table>
<thead>
<tr>
<th></th>
<th>$R$</th>
<th></th>
<th>$S$</th>
<th></th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\Phi$</td>
<td></td>
<td>A</td>
<td>$\Phi$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$x_1$</td>
<td></td>
<td>1</td>
<td>$y_1$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$x_2$</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

Tuples annotations modelled by semirings

- $(R \cup S) \Join_A T$ yields

<table>
<thead>
<tr>
<th></th>
<th>$(R \cup S) \Join_A T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\Phi$</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

- Aggregation on top of this table yields:

$$((x_1 + y_1) \cdot z_1) \otimes 17 + (x_2 \cdot z_2) \otimes 5$$

where the meaning of $+$ depends on the aggregation monoid.
Monoids, Semirings, **Semimodule**

**Semimodule**

- Algebraic framework introduced by Amstredamer et al. [2011]
- The algebraic structure combining semirings and monoids is called **semimodule**
- Generalisation of vector space. “Scalars”: tuple annotations, “Vectors”: aggregation values
- Semimodule expressions represent data values *conditioned* on tuple annotations

**Semiring and semimodule expressions are random variables**

- Semimodule: Random variable over aggregation domain
- Semiring expressions: ?
  - So far in probabilistic databases: *Boolean* random variable
  - However: 
    - is in general not large enough for aggregation; need larger semiring, for example natural numbers
Aggregation Needs Semirings Larger Than $\mathbb{B}$

<table>
<thead>
<tr>
<th>Item</th>
<th>$\Phi$</th>
<th>Item</th>
<th>$\Phi$</th>
<th>Item</th>
<th>Price</th>
<th>$\Phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$x_1$</td>
<td>1</td>
<td>$y_1$</td>
<td>1</td>
<td>17</td>
<td>$z_1$</td>
</tr>
<tr>
<td>2</td>
<td>$x_2$</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>$z_2$</td>
<td></td>
</tr>
</tbody>
</table>

- Query: $\text{SUM}_{\text{Price}} \left[ (\text{ProducerEU} \cup \text{ProducerUS}) \times_{\text{item}} \text{Products} \right]$ asking for total price of products sold by all producers

- Resulting expression: $\left( (x_1 + y_1) \cdot z_1 \right) \otimes 17 + (x_2 \cdot z_2) \otimes 5$

- Valuation $\nu : x_1, x_2, y_1, z_1, z_2 \mapsto \top$ yields $\top \otimes 17 + \top \otimes 5 = 22$

Arguably not the expected result
Aggregation Needs Semirings Larger Than $\mathbb{B}$

| ProducerEU | | ProducerUS | | Products |
|------------|------------|------------|------------|
| Item | $\Phi$ | Item | $\Phi$ | Item | Price | $\Phi$ |
| 1 | $x_1$ | 1 | $y_1$ | 1 | 17 | $z_1$ |
| 2 | $x_2$ | | | 2 | 5 | $z_2$ |

- **Query:** $\text{SUM}_{\text{Price}} \left[ (\text{ProducerEU} \cup \text{ProducerUS}) \otimes_{\text{Item}} \text{Products} \right]$ asking for total price of products sold by all producers

- **Resulting expression:** $((x_1 + y_1) \cdot z_1) \otimes 17 + (x_2 \cdot z_2) \otimes 5$

- **Valuation** $\nu : x_1, x_2, y_1, z_1, z_2 \mapsto \top$ yields $\top \otimes 17 + \top \otimes 5 = 22$
  Arguably not the expected result

- Boolean semiring is not large enough for SUM

- Better choice: Semiring $\mathbb{N}$. Identify $\bot \sim 0$, $\top \sim 1$.

- Valuation $\nu : x_1, x_2, y_1, z_1, z_2 \mapsto 1$ yields $( (1 + 1) \cdot 1 ) \otimes 17 + (1 \cdot 1) \otimes 5 = 2 \otimes 17 + 1 \otimes 5 = 39$. 

Outline

Motivation

Algebraic Foundations

Representation System

Query Evaluation
The pvc-tables Representation System

Ingredients for pvc-tables

- A set $X$ of variable symbols
- Tuples contain constants or semimodule expressions over $X$
- Every tuple is annotated with a semiring expression over $X$

Queries

- Query $Q$ maps pvc-table database $D$ to pvc-table $Q(D)$
- Annotations are propagated via query operators
- Expressions concisely encode probability distributions of answers

Properties of pvc-tables

- Polynomial overhead (Amsterdamer et al. [2011]):
  $$|Q(D)| \in \mathcal{O}(\text{poly}(|D|))$$
  (unlike pc-tables)
- Completeness: Every finite probability distribution over relations
  (with set or bag semantics) can be represented by pvc-tables
The pvc-tables Representation System

Semantics: Set vs Bag & Deterministic vs Probabilistic

Different choices for the semiring and the probability distributions of the annotation variables give rise to different database semantics.

<table>
<thead>
<tr>
<th>Database Semantics</th>
<th>Semiring</th>
<th>Probability Distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic</td>
<td>Set</td>
<td>$\mathbb{B}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$P_x[\top] = 1$ or $P_x[\bot] = 1$</td>
</tr>
<tr>
<td>Deterministic</td>
<td>Bag</td>
<td>$\mathbb{N}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\exists n \in \mathbb{N} : P_x[n] = 1$</td>
</tr>
<tr>
<td>Probabilistic</td>
<td>Set</td>
<td>$\mathbb{B}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$P_x[\top], P_x[\bot] \in [0, 1]$</td>
</tr>
<tr>
<td>Probabilistic</td>
<td>Bag</td>
<td>$\mathbb{N}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\forall n \in \mathbb{N} : P_x[n] \in [0, 1]$</td>
</tr>
</tbody>
</table>
Outline

Motivation

Algebraic Foundations

Representation System

Query Evaluation
Query Evaluation in pvc-tables (1)

Step 1: Construction of Expressions

Alongside (standard) query evaluation, compute annotations.

- Project, Union, Cartesian Product: Construction of semiring expressions (\( \cdot \) for joint, and \( + \) for alternative use of data)

- Aggregation (with grouping): Construct semimodule expressions
  
  \[
  \left( \sum_{\text{AGG}} \Phi \otimes \nu \right)
  \]

<table>
<thead>
<tr>
<th>R</th>
<th>A</th>
<th>B</th>
<th>( \Phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>1 ( x_1 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>2 ( x_2 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>3 ( x_3 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>4 ( x_4 )</td>
<td></td>
</tr>
</tbody>
</table>

select AGG(B) from R group by A

<table>
<thead>
<tr>
<th>pvc-table</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>a</td>
</tr>
<tr>
<td>b</td>
</tr>
</tbody>
</table>
Step 2: Probability Computation

Problem: Given a tuple, compute its probability distribution.

Idea: Tuple probability is equivalent to joint probability distribution of its semimodule expressions and annotation expression as obtained from evaluation step 1.

Approach: Compile expressions into a tractable form consisting of independent and mutually exclusive sub-expressions.
Consider semiring expression $\Phi = x + y$. If $x$, $y$ are independent random variables over $\mathbb{N}$, then the probability distribution of $\Phi$ is given by the convolution of $x$ and $y$.

If $x, y$ are in $\mathbb{N}$:  
\[
P_{x+y}[n] = \sum_{i,j \in \mathbb{N}, i+j=n} P_x[i]P_y[j]
\]
Consider semiring expression $\Phi = x + y$. If $x, y$ are independent random variables over $\mathbb{N}$, then the probability distribution of $\Phi$ is given by the convolution of $x$ and $y$.

If $x, y$ are in $\mathbb{N}$:  
\[
P_{x+y}[n] = \sum_{i,j \in \mathbb{N}} P_x[i]P_y[j] \quad \text{with} \quad i+j=n
\]

If $x, y$ are Boolean:  
\[
P_{x+y}[\bot] = \sum_{a,b \in \{\bot, \top\}} P_x[a]P_y[b] \quad \text{with} \quad a \lor b = \bot
\]
\[
P_{x+y}[\top] = \sum_{a,b \in \{\bot, \top\}} P_x[a]P_y[b] \quad \text{with} \quad a \lor b = \top
\]

\[
P_{x+y}[\top] = 1 - P_x[\bot]P_y[\bot]
\]
Compilation: Independent Decomposition

Consider semiring expression $\Phi = x + y$. If $x$, $y$ are independent random variables over $\mathbb{N}$, then the probability distribution of $\Phi$ is given by the convolution of $x$ and $y$.

If $x$, $y$ are in $\mathbb{N}$:  
$$P_{x+y}[n] = \sum_{i,j \in \mathbb{N}, i+j=n} P_x[i]P_y[j]$$

If $x$, $y$ are Boolean:  
$$P_{x+y}\bot = \sum_{a,b \in \{\bot, \top\} \atop a \lor b = \bot} P_x[a]P_y[b] = P_x[\bot]P_y[\bot]$$
$$P_{x+y}\top = \sum_{a,b \in \{\bot, \top\} \atop a \lor b = \top} P_x[a]P_y[b]$$
$$= P_x[\top]P_y[\top] + P_x[\bot]P_y[\top] + P_x[\top]P_y[\bot]$$
$$= 1 - P_x[\bot]P_y[\bot]$$
The applicability of convolution is not limited to “sums”; convolution is equally well defined for other binary operations:

Convolution for algebraic operations

- Semiring expressions: $\Phi \cdot \Psi$, $\Phi + \Psi$
- Semimodule expressions: $\alpha + \beta$
- Mixed semiring and semimodule expressions: $\Phi \otimes \alpha$
- Convolution is also applicable to comparisons of expressions, such as $\alpha \leq \beta$
Compilation: Mutually Exclusive Expressions

What if there are no independent sub-expressions?
Example: $\alpha = a(b + c) \otimes 10 + c \otimes 20$

Idea: Instantiate one of the variables to create mutually exclusive sub-expressions.

$$P(\alpha) = P_c[1] \cdot P(a(b + 1) \otimes 10 + 1 \otimes 20) +$$
$$P_c[2] \cdot P(a(b + 2) \otimes 10 + 2 \otimes 20) +$$
$$P_c[3] \cdot P(a(b + 3) \otimes 10 + 3 \otimes 20) +$$
$$\ldots$$

Need to consider all possible values of $c$ with non-zero probability. In particular: For Boolean variables, the above construction yields Shannon’s expansion.
Decomposition Trees (d-trees)

Decomposition gives rise to a tree whose nodes explain the decomposition steps taken. For example, $\cup$ for mutex decomposition, $\oplus$ for convolution w.r.t. $+$, $\otimes$ for convolution w.r.t. $\otimes$, etc.

Example: $\alpha = a(b + c) \otimes 10 + c \otimes 20$
The probability distribution $P_d$ of a d-tree $d$ whose nodes have probability distributions $p_1, \ldots, p_n$ can be computed in time $O(\prod |p_i|)$.

Specific polynomial time cases

- For MIN and MAX monoids combined with any semiring
- For SUM monoid: If monoid values and size of probability distributions of semiring expressions are bounded by constants
  - This subsumes COUNT aggregation
Further Applications of d-trees

- Approximate probability computation by partial expansion of d-tree (Olteanu et al. [2010], Fink et al. [2011])
- Sensitivity analysis and explanation of query results (Kanagal et al. [2011])
- Conditioning probabilistic databases (Koch and Olteanu [2008])
Tractable Queries via d-trees

**Tractability** for query evaluation on probabilistic databases is considered with respect to **data complexity**:

For which class of queries can probability distributions of query answers be computed in *polynomial-time data complexity* for any tuple-independent database?
Tractable Queries via d-trees

Tractability for query evaluation on probabilistic databases is considered with respect to data complexity:

For which class of queries can probability distributions of query answers be computed in \textit{polynomial-time data complexity} for any tuple-independent database?

- Syntactic characterisation of tractable queries with aggregates
  - There are known classes of tractable non-aggregate queries with polynomial-time d-tree compilation, e.g. hierarchical queries
  - Extend these classes by adding nested aggregation without breaking the tractable (e.g. hierarchical) property
Tractable Queries via d-trees

Example 1

$$\text{select R.A from R where R.B} = \left( \text{select MIN(S.B) from S where S.C = R.C} \right)$$

Tractable sub-queries without aggregation:

$$\text{select S.B from S where S.C = R.C}$$
Example 2

\[
\text{select 1 where } \left( \text{select } \text{MIN}(R.A) \text{ from } R \right) \leq \left( \text{select } \text{COUNT}(\ast) \text{ from } S,T \right) \text{ where } S.A=T.A
\]

Tractable sub-queries without aggregation:

select 1 where (select R.A from R)
select 1 from S,T where S.A=T.A

select 1 where (select R.A from R) \leq (select 1 from S,T where S.A=T.A)
End.


