Non-Standard-Datenbanken

Prof. Dr. Ralf Möller
Universität zu Lübeck
Institut für Informationssysteme
Acknowledgements:

This presentation is based on the following two presentations

10 Years of Probabilistic Querying – What Next?  
**Martin Theobald**  
University of Antwerp

Temporal Alignment  
**Anton Dignös**\(^1\) **Michael H. Böhlen**\(^1\) **Johann Gamper**\(^2\)  
\(^1\)University of Zürich, Switzerland  
\(^2\)Free University of Bozen-Bolzano, Italy
Recap: Probabilistic Databases

A probabilistic database $D^p$ (compactly) encodes a probability distribution over a finite set of deterministic database instances $D_i$.

**Special Cases:**

1. **Tuple-independent**

   - $D_1$: 0.42
     - `WorksAt(Sub, Obj)`
     - Jeff Stanford
     - Jeff Princeton

   - $D_2$: 0.18
     - `WorksAt(Sub, Obj)`
     - Jeff Stanford

   - $D_3$: 0.28
     - `WorksAt(Sub, Obj)`
     - Jeff Princeton

   - $D_4$: 0.12

2. **Block-independent**

   - **Note:** (I) and (II) are not equivalent!

**Query Semantics:** (“Marginal Probabilities”)

- Run query $Q$ against each instance $D_i$; for each answer tuple $t$, sum up the probabilities of all instances $D_i$ where $t$ is a result.
Probabilistic & Temporal Databases

A temporal-probabilistic database $D_{Tp}$ (compactly) encodes a probability distribution over a finite set of deterministic database instances $D_i$ at each time point of a finite time domain $\Omega_T$.

<table>
<thead>
<tr>
<th>BornIn(Sub,Obj)</th>
<th>T</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeNiro Greenwhich</td>
<td>[1943, 1944]</td>
<td>0.9</td>
</tr>
<tr>
<td>DeNiro Tribeca</td>
<td>[1998, 1999]</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wedding(Sub,Obj)</th>
<th>T</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeNiro Abbott</td>
<td>[1936, 1940]</td>
<td>0.3</td>
</tr>
<tr>
<td>DeNiro Abbott</td>
<td>[1976, 1977]</td>
<td>0.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Divorce(Sub,Obj)</th>
<th>T</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeNiro Abbott</td>
<td>[1988, 1989]</td>
<td>0.8</td>
</tr>
</tbody>
</table>

- **Sequenced Semantics & Snapshot Reducibility:**
  - Built-in semantics: reduce temporal-relational operators to their non-temporal counterparts at each snapshot (i.e., time point) of the database.
  - Coalesce/split tuples with consecutive time intervals based on their lineages.

- **Non-Sequenced Semantics**
  - Queries can freely manipulate timestamps just like regular attributes.
  - Single temporal operator $\leq^T$ supports all of Allen’s 13 temporal relations.
  - Deduplicate tuples with overlapping time intervals based on their lineages.

[Dignös, Gamper, Böhlen: SIGMOD’12]

marriedTo(x,y)[t_{b1},T_{max}) \iff wedding(x,y)[t_{b1},t_{e1}) \land \neg divorce(x,y)[t_{b2},t_{e2})

[Dylla, Miliaraki, Theobald: PVLDB’13]
Sequenced Semantics: Example

- **Input**: Employee $N$ works for department $D$ during time $T$.

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>$D$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$</td>
<td>Joe</td>
<td>DB</td>
<td>(Feb, Jul)</td>
</tr>
<tr>
<td>$r_2$</td>
<td>Ann</td>
<td>DB</td>
<td>(Feb, Sep)</td>
</tr>
<tr>
<td>$r_3$</td>
<td>Sam</td>
<td>AI</td>
<td>(May, Oct)</td>
</tr>
</tbody>
</table>

- **Query**: How did the average duration of contracts per department change?

- **Result**: Temporal Aggregation: $D \nu^T_{\text{AVG}}(\text{DUR}(T))(R)$

<table>
<thead>
<tr>
<th></th>
<th>AVG</th>
<th>$D$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_1$</td>
<td>6</td>
<td>DB</td>
<td>(Feb, Jul)</td>
</tr>
<tr>
<td>$z_2$</td>
<td>7</td>
<td>DB</td>
<td>(Jul, Sep)</td>
</tr>
<tr>
<td>$z_3$</td>
<td>5</td>
<td>AI</td>
<td>(May, Oct)</td>
</tr>
</tbody>
</table>

Timestamps must be adjusted for the result.

[Dignös, Gamper, Böhlen: SIGMOD'12]
Temporal Splitter / Snapshot Reduction

- Average duration of contracts per department: $D\forall^T_{AVG(DUR(T))}(R)$

![Diagram showing temporal splitting and snapshot reduction with example tuples and aggregation]

- One input tuple contributes to at most one result tuple per month.

[Dignös, Gamper, Böhlen: SIGMOD’12]
Non-Sequenced Semantics:

marriedTo(x,y)[t_{b1},T_{max}) ⇐ wedding(x,y)[t_{b1},t_{e1}) ∧ ¬divorce(x,y)[t_{b2},t_{e2})
marriedTo(x,y)[t_{b1},t_{e2}) ⇐ wedding(x,y)[t_{b1},t_{e1}) ∧ divorce(x,y)[t_{b2},t_{e2}) ∧ t_{e1} ≤ T t_{b2}
Derived Facts

\[ \text{teamMates}(\text{Beckham, Ronaldo, } T_3) \iff \text{playsFor}(\text{Beckham, Real, } T_1) \land \text{playsFor}(\text{Ronaldo, Real, } T_2) \land \text{overlaps}(T_1, T_2, T_3) \]

Base Facts

\[ \text{playsFor}(\text{Beckham, Real, } T_1) \]
\[ \text{playsFor}(\text{Ronaldo, Real, } T_2) \]

Example using the Allen predicate \textit{overlaps}
Inference in Probabilistic-Temporal Databases

[Derived Facts]

\( \text{playsFor}(\text{Beckham, Real, } T_1) \)  
\( \text{playsFor}(\text{Ronaldo, Real, } T_2) \)  
\( \text{teamMates}(\text{Beckham, Ronaldo, } T_4) \)  
\( \text{teamMates}(\text{Beckham, Zidane, } T_5) \)  
\( \text{teamMates}(\text{Ronaldo, Zidane, } T_6) \)  

[Base Facts]

Non-independent

Independent

[Inference in Probabilistic-Temporal Databases]

[Wang, Yahya, Theobald: MUD’10; Dylla, Miliaraki, Theobald: PVLDB’13]
Inference in Probabilistic-Temporal Databases

[Wang,Yahya,Theobald: MUD’10; Dylla,Miliaraki,Theobald: PVLDB’13]

Derived facts stored in views

- \text{teamMates}(Beckham, Ronaldo, T_4)
- \text{teamMates}(Beckham, Zidane, T_5)
- \text{teamMates}(Ronaldo, Zidane, T_6)

Closed and complete representation model (incl. lineage)

Temporal alignment is polyn. in the number of input intervals

Confidence computation per interval remains \#P-hard

In general requires Monte Carlo approximations (Luby-Karp for DNF, MCMC-style sampling), decompositions, or top-k pruning

Need Lineage!
Lineage & Possible Worlds

1) Deductive Grounding
   - Dependency graph of the query
   - Trace lineage of individual query answers

2) Lineage DAG (not in CNF), consisting of
   - Grounded soft & hard views
   - Probabilistic base facts

3) Probabilistic Inference
   → Compute marginals:
   - $P(Q)$: sum up the probabilities of all possible worlds that entail the query answers’ lineage
   - $P(Q|H)$: drop “impossible worlds”

Query

graduatedFrom(Surajit, y)

$0.7 \times (1 - 0.888) = 0.078$

$\neg \text{graduatedFrom(Surajit, Princeton)}$ $\land \neg (B \lor (C \land D))$

$0.8 \times 0.9 = 0.72$

$\text{graduatedFrom(Surajit, Princeton)}[0.7]$

$(1 - 0.7) \times 0.888 = 0.266$

$\neg \text{graduatedFrom(Surajit, Stanford)}$

$\neg A \lor (B \lor (C \land D))$

$1 - (1 - 0.72) \times (1 - 0.6) = 0.888$

$\text{graduatedFrom(Surajit, Stanford)}[0.6]$

$\text{hasAdvisor(Surajit, Jeff)}[0.8]$

$\text{worksAt(Jeff, Stanford)}[0.9]$

[Das Sarma, Theobald, Widom: ICDE'08
Dylla, Miliaraki, Theobald: ICDE'13]
[Das Sarma, Theobald, Widom: ICDE’08]

[Wang, Yahya, Theobald: MUD’10]

[Dignös, Gamper, Böhlen: SIGMOD’12]

[Dylla, Miliaraki, Theobald: PVLDB’13]
Historical Facts vs. Future Facts

- Processing uncertain historical data

\[
\text{marriedTo}(x,y)[t_{b1}, t_{e2}) \iff \text{wedding}(x,y)[t_{b1}, t_{e1}) \land \neg \text{divorce}(x,y)[t_{b2}, t_{e2}) \\
\text{marriedTo}(x,y)[t_{b1}, t_{e2}) \iff \text{wedding}(x,y)[t_{b1}, t_{e1}) \land \text{divorce}(x,y)[t_{b2}, t_{e2}) \land t_{e1} \leq T_{t_{b2}}
\]

- Estimating probabilities of future facts (increasing \(T_{max}\))?

\[
\text{marriedTo}(x,y)[t_{b1}, T_{max}) \iff \text{wedding}(x,y)[t_{b1}, t_{e1}) \land \neg \text{divorce}(x,y)[t_{b2}, t_{e2}) \\
\text{marriedTo}(x,y)[t_{b1}, t_{e2}) \iff \text{wedding}(x,y)[t_{b1}, t_{e1}) \land \text{divorce}(x,y)[t_{b2}, t_{e2}) \land t_{e1} \leq T_{t_{b2}}
\]