Representing and Querying Correlated Tuples in Probabilistic Databases

Presented by Kun Li
kli@cise.ufl.edu
Introduction

- Abundance of uncertain data.

- Numerous approaches proposed to handle uncertainty.

- However, most models make assumptions about data uncertainty that restricts applicability.

**What we need?**

- a model of uncertainty that can capture correlations

- simple and intuitive semantics that are readily understood and defines precise answers to every query
Motivation for correlated data

Applications:

- “Dirty” databases [CP87, DS96, AFM06]: Arise while trying to integrate data from various sources.

- Sensor Networks Often show strong spatial correlations, e.g., nearby sensors report similar values.
Outline

1. Motivation.

1. Probabilistic databases with correlations.

1. Query evaluation.

1. Qualitative comparison with other approaches.

1. Experiments.

1. Conclusion and future work
Independent tuple-based probabilistic databases

\[
D^p
\]

\[
\begin{array}{|c|c|}
\hline
S^p & \text{prob} \\
\hline
s_1 & A \quad B \\
\hline
m & 0.6 \\
n & 0.5 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
T^p & \text{prob} \\
\hline
t_1 & C \quad D \\
\hline
1 & p \\
\hline
\end{array}
\]

\[
pwd(D^p)
\]

\[
\begin{array}{|c|c|}
\hline
\text{instance} & \text{probability} \\
\hline
d_1 = \{s_1, s_2, t_1\} & 0.12 \\
d_2 = \{s_1, s_2\} & 0.18 \\
d_3 = \{s_1, t_1\} & 0.12 \\
d_4 = \{s_1\} & 0.18 \\
d_5 = \{s_2, t_1\} & 0.08 \\
d_6 = \{s_2\} & 0.12 \\
d_7 = \{t_1\} & 0.08 \\
d_8 = \emptyset & 0.12 \\
\hline
\end{array}
\]

\[
\text{Evaluation}
\]

\[
\begin{array}{|c|}
\hline
\text{query result} \\
\hline
\{p\} \\
\emptyset \\
\{p\} \\
\emptyset \\
\{p\} \\
\emptyset \\
\emptyset \\
\emptyset \\
\hline
\end{array}
\]

\[
\text{Result}
\]

\[
\begin{array}{|c|}
\hline
D & \text{prob} \\
p & \text{prob}(d_1) + \text{prob}(d_3) + \text{prob}(d_5) = 0.32 \\
\emptyset & \text{prob}(d_2) + \text{prob}(d_4) + \text{prob}(d_6) + \text{prob}(d_7) + \text{prob}(d_8) = 0.68 \\
\hline
\end{array}
\]
How to represent correlations

- Associate with each tuple in the probabilistic database a random variable.

- Define factors on (sub)sets of tuple-based random variables to encode correlations.

- The probability of an instantiation of the database is given by the product of all the factors.
Tuple Correlations

1. *ind.*: where $s_1$, $s_2$, and $t_1$ are independent of each other.

2. *implies*: presence of $t_1$ implies absence of $s_1$ and $s_2$ ($t_1 \Rightarrow \neg s_1 \land \neg s_2$).

3. *mutual exclusivity* (*mut. ex.*): $t_1 \Rightarrow \neg s_1$ and $s_1 \Rightarrow \neg t_1$.

4. *nxor*: high positive correlation between $t_1$ and $s_1$, presence/absence of one almost certainly implies the presence/absence of the other.
Representing correlations: Mutual Exclusivity Example

\[ D^p \]

\[
\begin{array}{c|c|c|c}
S^p & \text{prob} & \text{implies} \\
\hline
A & 0.6 & 0 \\
B & 0.5 & 0.5 \\
\end{array}
\]

\[
\begin{array}{c|c}
T^p & \text{prob} \\
\hline
C & 0.4 \\
D & 0.4 \\
\end{array}
\]

\[ Pr(X_{DP}) = f_{s_2}^{ind}(X_{s_2}) \]
\[ f_{t_1,s_1}^{mutex}(X_{t_1},X_{s_1}) \]

\[
\begin{array}{ccc}
X_{t_1} & X_{s_1} & f_{t_1,s_1}^{mutex} \\
0 & 0 & 0 \\
0 & 1 & 0.6 \\
1 & 0 & 0.4 \\
1 & 1 & 0 \\
\end{array}
\]
Difference in query result

\[ \prod_D(S^p \uplus_{B=C} T^p) \]

<table>
<thead>
<tr>
<th>$D^p$</th>
<th>$S^p$</th>
<th>$T^p$</th>
<th>prob</th>
<th>instance</th>
<th>ind.</th>
<th>implies</th>
<th>mut. ex.</th>
<th>nxor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>prob</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>D</td>
<td>prob</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_1$</td>
<td>m</td>
<td>1</td>
<td>0.6</td>
<td>$d_1 = {s_1, s_2, t_1}$</td>
<td>0.12</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>1</td>
<td>0.5</td>
<td>$d_2 = {s_1, s_2}$</td>
<td>0.18</td>
<td>0.5</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>$s_2$</td>
<td></td>
<td></td>
<td></td>
<td>$d_3 = {s_1, t_1}$</td>
<td>0.12</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$d_4 = {s_1}$</td>
<td>0.18</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$d_5 = {s_2, t_1}$</td>
<td>0.08</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$d_6 = {s_2}$</td>
<td>0.12</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$d_7 = {t_1}$</td>
<td>0.08</td>
<td>0.4</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$d_8 = \emptyset$</td>
<td>0.12</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D$</th>
<th>ind.</th>
<th>implies</th>
<th>mut. ex.</th>
<th>nxor</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.32</td>
<td>0.00</td>
<td>0.20</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Note: The table entries and formulas are placeholders; actual values should be provided based on the context.
Probabilistic Graphical Models and Factored Representations

$$Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3) =$$

$$f_1(X_1 = x_1)f_{12}(X_1 = x_1, X_2 = x_2)f_{23}(X_2 = x_2, X_3 = x_3)$$

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$f_1$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$f_{12}$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$f_{23}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>0</td>
<td>1</td>
<td>0.1</td>
<td>0</td>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>0.7</td>
</tr>
</tbody>
</table>

(i)

$$
\begin{array}{|c|c|c|}
\hline
x_1 & x_2 & x_3 \\
\hline
0 & 0 & 0 & 0.378 \\
0 & 0 & 1 & 0.162 \\
0 & 1 & 0 & 0.018 \\
0 & 1 & 1 & 0.042 \\
1 & 0 & 0 & 0.028 \\
1 & 0 & 1 & 0.012 \\
1 & 1 & 0 & 0.108 \\
1 & 1 & 1 & 0.252 \\
\hline
\end{array}
$$

(ii)

(iii)
Representing dependency

- For each possible tuple t, let boolean-valued random variable $X_t$ represent its presence (1) or absence (0) in the database.
- Joint distribution of vector X represents all possible worlds with probabilities.
- Worst case: represent the full joint distribution, e.g.:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$Pr$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.378</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.018</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.028</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.012</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.108</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.252</td>
</tr>
</tbody>
</table>
A more efficient representation

- Idea from probabilistic graphical model.
- Use factored representation, where each factor $f_i$

$$Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3) = f_1(x_1) f_{12}(x_1, x_2) f_{23}(x_2, x_3)$$

$$Pr(X = x) = \prod_{i=1}^{m} f_i(X_i = x_i)$$

(i)

(ii)

(iii)
Example revisited

**Implies:** \((t_1 \Rightarrow \neg s_1 \land \neg s_2)\).

\[
Pr(X_{DP}) = f_{t_1}^{ind}(X_{t_1}) f_{t_1,s_1}^{implies}(X_{t_1}, X_{s_1})
\]

\[
f_{t_1,s_1}^{implies}(X_{t_1}, X_{s_1})
\]

<table>
<thead>
<tr>
<th>(X_{t_1})</th>
<th>(X_{s_1})</th>
<th>(f_{t_1,s_1}^{implies})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
f_{t_1,s_1}^{mutext}(X_{t_1}, X_{s_1})
\]

<table>
<thead>
<tr>
<th>(X_{t_1})</th>
<th>(X_{s_1})</th>
<th>(f_{t_1,s_1}^{mutext})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
Pr(X_{DP}) = f_{s_2}^{ind}(X_{s_2})
\]

\[
f_{t_1,s_1}^{nxor}(X_{t_1}, X_{s_1})
\]

<table>
<thead>
<tr>
<th>(X_{t_1})</th>
<th>(X_{s_1})</th>
<th>(f_{t_1,s_1}^{nxor})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Query Evaluation: Details

- Basic idea: build up the joint distribution of base and derived tuples as we go.
- Each operator introduces a set of new (output) tuples, and a factor connecting them to existing (input) tuples.
Representing Probabilistic Relations
From factors to tuple probabilities

- Cast as an inference problem in graphical models
- Example: variable elimination
  - Given the factors, what is \( \Pr(x_3) \)?

\[
\Pr(x_3) = \sum_{x_2, x_1} f_1(x_1)f_{12}(x_1,x_2)f_{23}(x_2,x_3)
\]

\[
= \sum_{x_2} f_{23}(x_2,x_3) \sum_{x_1} f_1(x_1)f_{12}(x_1,x_2)
\]

\[
= \sum_{x_2} f_{23}(x_2,x_3) \mu_1(x_2)
\]

\[
= \sum_{x_2} f_{23}(x_2,x_3) \mu_1(x_2) \mu_2(x_3)
\]
Need for Modeling Dependencies

- Exact data: PUBS(PID,Title) AUTHS(PID,Name).
- Query with approximate predicate

\[ \Pi_{\text{Title}}(\sigma_{\text{Name}\approx x}(\text{AUTHS}) \bowtie \sigma_{\text{Title}\approx y}(\text{PUBS})) \]

<table>
<thead>
<tr>
<th>Title</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reinforcement learning with hidden states (by L. Lin, T. Mitchell)</td>
<td>(i) MUTE-B results at ( \sigma = 10, 50, 100 )</td>
<td></td>
</tr>
<tr>
<td>Feudal Reinforcement Learning (by C. Atkeson, P. Dayan, ...)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reasoning (by C. Bereiter, M. Scardamalia)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| ...

<table>
<thead>
<tr>
<th>Title</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reinforcement learning with hidden states (by L. Lin, T. Mitchell)</td>
<td>(ii) IND-B results at ( \sigma = 10 )</td>
<td></td>
</tr>
<tr>
<td>Feudal Reinforcement Learning (by C. Atkeson, P. Dayan, ...)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reasoning (by C. Bereiter, M. Scardamalia)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| ...

<table>
<thead>
<tr>
<th>Title</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Feudal Reinforcement Learning (by C. Atkeson, P. Dayan, ...)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision making and problem solving (G. Dantzig, R. Hogarth, ...)</td>
<td>(iii) IND-B results at ( \sigma = 50 )</td>
<td></td>
</tr>
<tr>
<td>Multimodal Learning Interfaces (by U. Bub, R. Houghton, ...)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| ...

<table>
<thead>
<tr>
<th>Title</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision making and problem solving (G. Dantzig, R. Hogarth, ...)</td>
<td>(iv) IND-B results at ( \sigma = 100 )</td>
<td></td>
</tr>
<tr>
<td>HERMES: A heterogeneous reasoning and mediator system (by S. Adali, A. Brink, ...)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Induction and reasoning from cases (by K. Althoff, E. Auriol, ...)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| ...


Performance Evaluation

- Small TPC-H benchmark (10MB)

Bare Query: database operations using our Java implementation.

Full Query: Bare Query Time + the probabilistic computations time
Conclusions and Future Work

Conclusions:
✓ Introduced probabilistic databases with correlated tuples.
✓ Utilized ideas from probabilistic graphical models to represent such correlations.
✓ Cast the query evaluation problem as an inference problem

Future Work:
✓ exact query evaluation over large datasets exhibiting complex correlations may not always be feasible. develop approximate query evaluation techniques.
✓ develop disk-based query evaluation algorithms.
Thanks! Any Questions?