Probabilistic Databases with Correlated Tuples

Prithviraj Sen   Amol Deshpande   Lise Getoor

Department of Computer Science
University of Maryland, College Park.

Uncertain/Probabilistic Databases, 2006

Introduction

- Abundance of uncertain data.
- Numerous approaches proposed to handle uncertainty [BP82, IWL84, FR97, LLRS97, CKP03, DS04, Wid05, CBL06].
- However, most models make assumptions about data uncertainty that restricts applicability.

Need for a Database model with

- a model of uncertainty that can capture correlations
- simple and intuitive semantics that is readily understood and defines precise answers to every query

Motivations for Correlated Data

Applications:

- "Dirty" databases [CP87, DS96, AFM06]: Arises while trying to integrate data from various sources.
- Sensor Networks [DGM$^+$04]: Often shows strong spatial correlations, e.g., nearby sensors report similiar values.
- More applications: Pervasive computing, approximate string matching in DB systems [DS04] etc.

Additional motivation:

- Correlated tuples arise while evaluating queries [DS04].

Representing Correlated Tuples: Ingredients

Which theory to use?

Probability theory , Dempster-Schafer theory, Fuzzy Logic, Logic ....

At what level do we represent uncertainty?

Tuple level , Attribute level.

Approach to representing correlations?

Probabilistic Graphical Models
(include as special cases: Bayesian networks and Markov networks)

Review: Independent Tuple-based Probabilistic Databases Possible World Semantics [DS04]

Example borrowed from [DS04]:

S

	A	B	*prob*
s_1	m	1	0.6
s_2	n	1	0.5

T

	C	D	*prob*
t_1	1	p	0.4

possible worlds

instance	probability
$\{s_1, s_2, t_1\}$	0.12
$\{s_1, s_2\}$	0.18
$\{s_1, t_1\}$	0.12
$\{s_1\}$	0.18
$\{s_2, t_1\}$	0.08
$\{s_2\}$	0.12
$\{t_1\}$	0.08
$\emptyset$	0.12

Representing Correlations: Basic Ideas

Two simple concepts borrowed from probabilistic graphical models literature [Pea88]:

Tuple-based Random Variables

Associate every tuple t with a boolean valued random variable X_t.

Factors

- $f(\mathbf{X})$ is a function of a (small) set of random variables $\mathbf{X}$.
- $0 \leq f(\mathbf{X}) \leq 1$

Representing Correlations: Basic Ideas - Contd.

- 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.

Representing Correlations: Mutual Exclusivity example

Suppose we want to represent mutual exclusivity between tuples s_1 and t_1. In particular, let us try to represent the possible worlds:

S

	A	B	*prob*
s_1	m	1	0.6
s_2	n	1	0.5

T

	C	D	*prob*
t_1	1	p	0.4

possible worlds

instance	probability
$\{s_1, s_2, t_1\}$	0
$\{s_1, s_2\}$	0.3
$\{s_1, t_1\}$	0
$\{s_1\}$	0.3
$\{s_2, t_1\}$	0.2
$\{s_2\}$	0
$\{t_1\}$	0.2
$\emptyset$	0

X_{t_1}	X_{s_1}	f_1
0	0	0
0	1	0.6
1	0	0.4
1	1	0

X_{s_2}	f_2
0	0.5
1	0.5

Suppose we want to represent positive correlation between t_1 and s_1. In particular, let us try to represent the possible worlds:

S

	A	B
s_1	m	1
s_2	n	1

T

	C	D
t_1	1	p

possible worlds

instance	probability
$\{s_1, s_2, t_1\}$	0.2
$\{s_1, s_2\}$	0.1
$\{s_1, t_1\}$	0.2
$\{s_1\}$	0.1
$\{s_2, t_1\}$	0
$\{s_2\}$	0.2
$\{t_1\}$	0
$\emptyset$	0.2

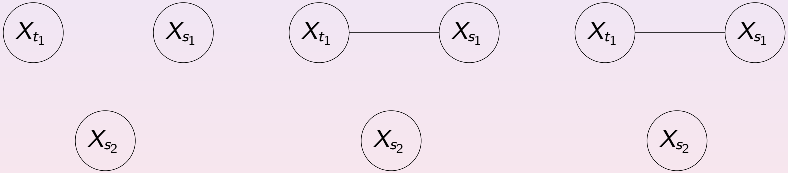
X_{t_1}	X_{s_1}	f_1
0	0	0.4
0	1	0.2
1	0	0
1	1	0.4

X_{s_2}	f_2
0	0.5
1	0.5

Probabilistic Graphical Model representation

Definition

A *probabilistic graphical model* is graph whose nodes represent random variables and edges represent correlations [Pea88].



Complete Ind. Example | Mutual Exclusivity Example | Positive Correlation Example

Query Evaluation: Basic Ideas

- Treat intermediate tuples as regular tuples.
- Carefully represent correlations between intermediate tuples, base tuples and result tuples to construct a probabilistic graphical model.
- Cast the probability computations resulting from query evaluation to *inference* in probabilistic graphical models.

Query Evaluation Example

Compute $\prod_{\mathbf{D}}(S \bowtie_{\mathbf{B}=\mathbf{C}} T)$

S:

	A	B
s_1	m	1
s_2	n	1

f_{s_1}, f_{s_2}

T:

	C	D
t_1	1	p

f_{t_1}

$S \bowtie_{\mathbf{B}=\mathbf{C}} T$ →

$f^{AND}_{i_1,s_1,t_1}, f^{AND}_{i_2,s_2,t_1}$

	A	B	C	D
i_1	m	1	1	p
i_2	n	1	1	p

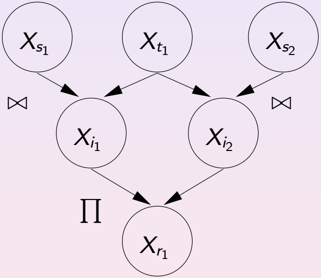
$\prod_{\mathbf{D}}(S \bowtie_{\mathbf{B}=\mathbf{C}} T)$ ↓

	D
r_1	p

$f^{OR}_{r_1,i_1,i_2}$

Query Evaluation Example: Prob. Graphical Model

- Query evaluation problem in Prob. Databases: Compute the probability of the result tuple summed over all possible worlds of the database [DS04].



- Equivalent problem in prob. graph. models: *marginal probability* computation.
- Thus we can use inference algorithms (e.g., *VE* [ZP94]).

Comparison with other probability computation approaches

Compute $\prod_{\{\}}(L \bowtie J \bowtie R)$:

L

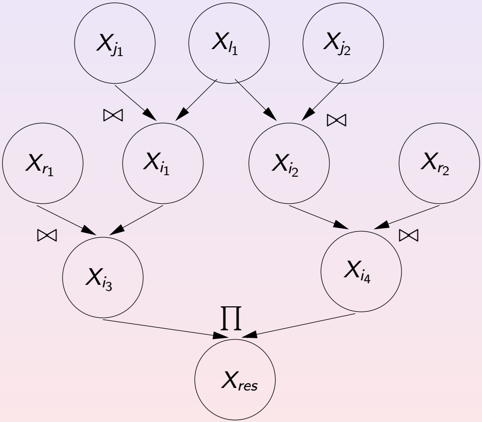
	A	B
l_1	m	1

J

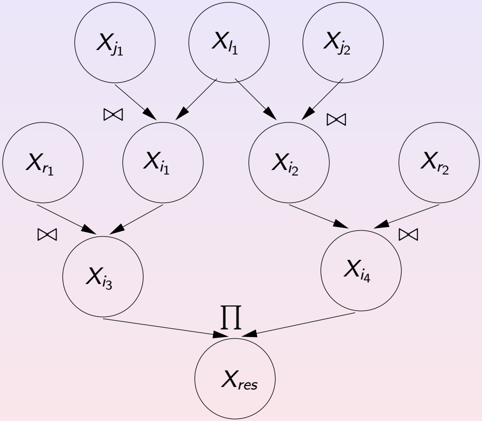
	B	C
j_1	1	p
j_2	1	q

R

	C	D
r_1	p	1
r_2	q	1

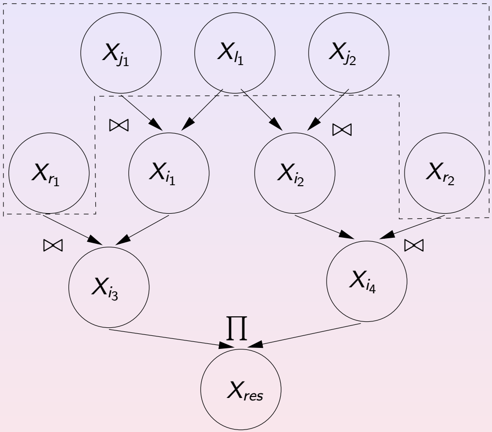


Comparison with Extensional Semantics [FR97, DS04]



- Is not guaranteed to match possible world semantics.
- Safe plans [DS04] return tree-structured graphical models.

Comparison with Intensional Semantics [FR97, DS04]



$$(X_{r_1} \wedge X_{j_1} \wedge X_{l_1}) \vee (X_{l_1} \wedge X_{j_2} \wedge X_{r_2})$$

- Work with a subgraph of the graph. model.
- Inference algorithms can exploit graph structure better.

Conclusion & Future Work

- Introduced Probabilistic Databases with correlated tuples.
- Borrowed ideas from Probabilistic Graphical Models to represent such correlations.
- Cast the query evaluation problem as an inference problem.
- Future Work:
    - Ways to restructure graphical model to speed up inference.
    - Share/reuse computation to speed up inference.
    - Explore the use of approximate inference methods [GRS96, JGJS99].

Thank You.

References

- Periklis Andritsos, Ariel Fuxman, and Renee J. Miller.
  Clean answers over dirty databases.
  In *International Conference on Data Engineering*, 2006.
- Bill P. Buckles and Frederick E. Petry.
  A fuzzy model for relational databases.
  International Journal of Fuzzy Sets and Systems, 1982.
- Sunil Choenni, Henk Ernst Blok, and Erik Leertouwer.
  Handling uncertainty and ignorance in databases: A rule to combine dependent data.
  In *Database Systems for Advanced Applications*, 2006.
- Reynold Cheng, Dmitri Kalashnikov, and Sunil Prabhakar.
  Evaluating probabilistic queries over imprecise data.
  In *International Conference on Management of Data.*, 2003.
- Roger Cavallo and Michael Pittarelli.
  The theory of probabilistic databases.

In *International Conference on Very Large Data Bases*, 1987.

- Amol Deshpande, Carlos Guestrin, Sam Madden, Joseph M. Hellerstein, and Wei Hong.
  Model-driven data acquisition in sensor networks.
  In *International Conference on Very Large Data Bases*, 2004.

- Debabrata Dey and Sumit Sarkar.
  A probabilistic relational model and algebra.
  ACM Transactions on Database Systems., 1996.

- Nilesh Dalvi and Dan Suciu.
  Efficient query evaluation on probabilistic databases.
  In *International Conference on Very Large Data Bases*, 2004.

- Norbert Fuhr and Thomas Rolleke.
  A probabilistic relational algebra for the integration of information retrieval and database systems.
  ACM Transactions on Information Systems, 1997.

- Walter R. Gilks, Sylvia Richardson, and David J. Spiegelhalter.

Markov Chain Monte Carlo in Practice.
Chapman & Hall, 1996.

Tomasz Imielinski and Jr. Witold Lipski.
Incomplete information in relational databases.
Journal of the ACM., 1984.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul.
An introduction to variational methods for graphical models.
Machine Learning, 1999.

Laks V. S. Lakshmanan, Nicola Leone, Robert Ross, and V. S. Subrahmanian.
Probview: a flexible probabilistic database system.
ACM Transactions on Database Systems., 1997.

Judaea Pearl.
Probabilistic Reasoning in Intelligent Systems.
Morgan Kaufmann, 1988.

Jennifer Widom.

Trio: A system for integrated management of data, accuracy, and lineage.
In *Proceedings of the Biennial Conference on Innovative Data Systems Research*, 2005.

Nevin Lianwen Zhang and David Poole.
A simple approach to bayesian network computations.
In *Canadian Conference on Artificial Intelligence*, 1994.