Efficient Query Computing for Uncertain Possibilistic Databases with Provenance

Angelos Vasilakopoulos Verena Kantere

National Technical University of Athens

Cyprus University of Technology

June 20, 2011

The problem we investigate:

How to compute answers to queries for uncertain data with attached "confidence values"?

Problem 1

Existing models for uncertain data (e.g., possibility theory) are not closed for SPJ queries.

Problem 2

Computing probabilities of SPJ answers in models that combine uncertainty and provenance is a problem with #P complexity.

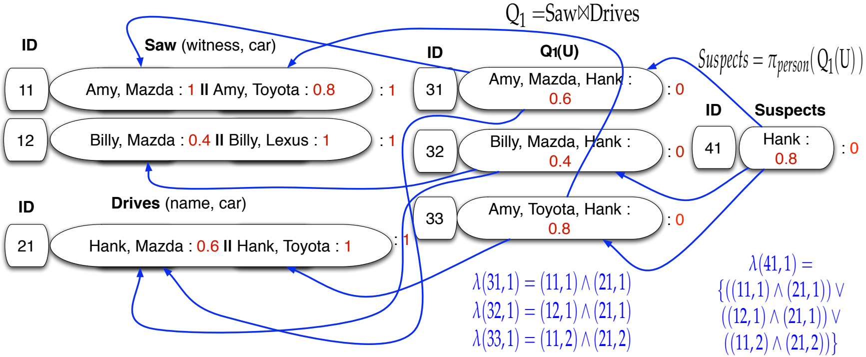
We propose:

A data model that combines uncertainty, provenance and possibilities.

Benefits of the proposed model:

- Closed for SPJ queries
- Computing possibilities of SPJ answers has PTIME complexity.

Motivating Example



Possible Worlds:

W1

ID	**Saw** (witness, car)	
11, 1	Amy, Mazda : 1	: 0.2
12, 1	Billy, Mazda : 0.4	: 0

ID	**Drives** (name, car)	
21, 1	Hank, Mazda : 0.6	: 0

ID	$Q_1(U)$	
31, 1	Amy, Mazda, Hank	: 0
32, 1	Billy, Mazda, Hank	: 0

ID	**Suspects**	
41	Hank	: 0

Π(W1)= 0.4 N(W1)= 0

W2

ID	**Saw** (witness, car)	
11, 1	Amy, Mazda : 1	: 0.2
12, 2	Billy, Lexus : 1	: 0.6

ID	**Drives** (name, car)	
21, 1	Hank, Mazda : 0.6	: 0

ID	$Q_1(U)$	
31, 1	Amy, Mazda, Hank	: 0

ID	**Suspects**	
41	Hank	: 0

Π(W2)= 0.6 N(W2)= 0

W3

ID	**Saw** (witness, car)	
11, 2	Amy, Toyota : 0.8	: 0
12, 1	Billy, Mazda : 0.4	: 0

ID	**Drives** (name, car)	
21, 1	Hank, Mazda : 0.6	: 0

ID	$Q_1(U)$	
32, 1	Billy, Mazda, Hank	: 0

ID	**Suspects**	
41	Hank	: 0

Π(W3)= 0.4 N(W3)= 0

W4

ID	**Saw** (witness, car)	
11, 2	Amy,Toyota : 0.8	: 0
12, 1	Billy, Mazda : 0.4	: 0

ID	**Drives** (name, car)	
21, 2	Hank,Toyota : 1	: 0.4

ID	$Q_1(U)$	
33, 1	Amy,Toyota, Hank	: 0

ID	**Suspects**	
41	Hank	: 0

Π(W4)= 0.4 N(W4)= 0

W5

ID	**Saw** (witness, car)	
11, 2	Amy, Toyota : 0.8	: 0
12, 2	Billy, Lexus : 1	: 0.6

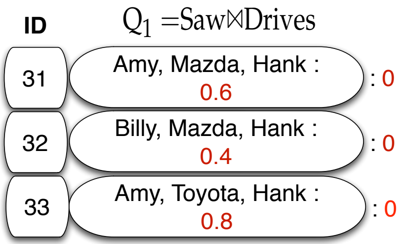
ID	**Drives** (name, car)	
21, 2	Hank, Toyota : 1	: 0.4

ID	$Q_1(U)$	
33, 1	Amy,Toyota, Hank	: 0

ID	**Suspects**	
41	Hank	: 0

Π(W5)= 0.8 N(W5)= 0

Possibilistic Databases are not closed



- Tuples 31 and 33 cannot coexist in any possible world.
- Adding provenance (lineage) makes possibilistic uncertain data model closed for SPJ queries.

A. Das Sarma et al. “Representing Uncertain Data: Models, Properties, and Algorithms”. VLDB Journal, October, 2009.
P. Bosc and O. Pivert. “About PSJ queries addressed to possibilistic relational databases”. IEEE T. Fuzzy Systems, 2005.

Computing probabilities is intractable

$$Suspects = \pi_{person}\left(Q_1(U)\right)$$

ID	Suspects
41	Hank

$$\lambda(41,1) = \{((11,1) \wedge (21,1)) \vee ((12,1) \wedge (21,1)) \vee ((11,2) \wedge (21,2))\}$$

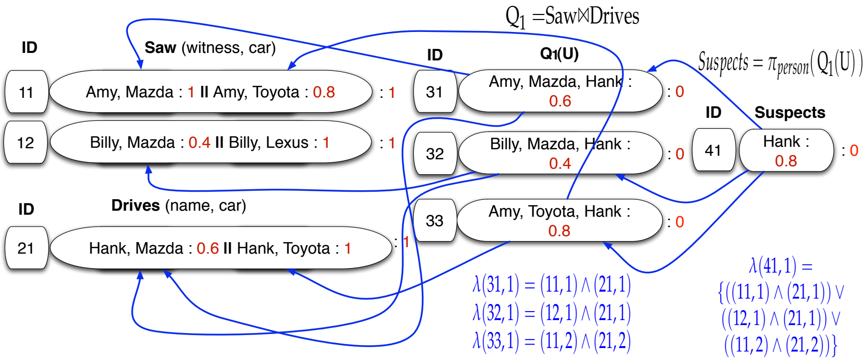
: Computing probabilities has #P complexity

N. Dalvi and D. Suciu. "Efficient query evaluation on probabilistic databases." In VLDB 2004.

Axioms of Possibilistic theory:

- $\Pi(X) = 1$
- $\Pi(\emptyset) = 0$
- $\Pi(E_1 \cup E_2) = \max(\Pi(E_1), \Pi(E_2))$
- $\Pi(E_1 \cap E_2) \leq \min(\Pi(E_1), \Pi(E_2))$
- $\Pi(E_1 \cap E_2) = \min(\Pi(E_1), \Pi(E_2))$
  (for not-interactive events)
- $\max\{\Pi(E),\ \Pi(\bar{E})\} = 1$
- $N(E) = 1 - \Pi(\bar{E})$

Running Example



Possible Worlds:

W1

ID	Saw (witness, car)	
11, 1	Amy, Mazda : 1	: 0.2
12, 1	Billy, Mazda : 0.4	: 0

ID	Drives (name, car)	
21, 1	Hank, Mazda : 0.6	: 0

ID	$Q_1(U)$	
31, 1	Amy, Mazda, Hank	: 0
32, 1	Billy, Mazda, Hank	: 0

ID	Suspects	
41	Hank	: 0

Π(W1)= 0.4 N(W1)= 0

W2

ID	Saw (witness, car)	
11, 1	Amy, Mazda : 1	: 0.2
12, 2	Billy, Lexus : 1	: 0.6

ID	Drives (name, car)	
21, 1	Hank, Mazda : 0.6	: 0

ID	$Q_1(U)$	
31, 1	Amy, Mazda, Hank	: 0

ID	Suspects	
41	Hank	: 0

Π(W2)= 0.6 N(W2)= 0

W3

ID	Saw (witness, car)	
11, 2	Amy, Toyota : 0.8	: 0
12, 1	Billy, Mazda : 0.4	: 0

ID	Drives (name, car)	
21, 1	Hank, Mazda : 0.6	: 0

ID	$Q_1(U)$	
32, 1	Billy, Mazda, Hank	: 0

ID	Suspects	
41	Hank	: 0

Π(W3)= 0.4 N(W3)= 0

W4

ID	Saw (witness, car)	
11, 2	Amy,Toyota : 0.8	: 0
12, 1	Billy, Mazda : 0.4	: 0

ID	Drives (name, car)	
21, 2	Hank,Toyota : 1	: 0.4

ID	$Q_1(U)$	
33, 1	Amy,Toyota, Hank	: 0

ID	Suspects	
41	Hank	: 0

Π(W4)= 0.4 N(W4)= 0

W5

ID	Saw (witness, car)	
11, 2	Amy, Toyota : 0.8	: 0
12, 2	Billy, Lexus : 1	: 0.6

ID	Drives (name, car)	
21, 2	Hank, Toyota : 1	: 0.4

ID	$Q_1(U)$	
33, 1	Amy,Toyota, Hank	: 0

ID	Suspects	
41	Hank	: 0

Π(W5)= 0.8 N(W5)= 0

Ongoing Work:

- Extend the query language to extentions of conjunctive queries.
- Find for which class of query languages the problem remains in polynomial time.
- Find for which class of query languages the problem becomes intractable.
- Study complexity of new query languages that can query over uncertainty and provenance.

Thank you