
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 Dignös¹ Michael H. Böhlen¹ Johann Gamper²

¹University of Zürich, Switzerland

²Free University of Bozen-Bolzano, Italy

Recap: Probabilistic Databases

A probabilistic database \mathbf{D}^p (compactly) encodes a probability distribution over a finite set of deterministic database instances \mathbf{D}_i .

$D_1: 0.42$

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

$D_4: 0.12$

WorksAt(Sub, Obj)	
-------------------	--

▶ Special Cases:

(I) \mathbf{D}^p tuple-independent

WorksAt(Sub, Obj)		p
Jeff	Stanford	0.6
Jeff	Princeton	0.7

(II) \mathbf{D}^p block-independent

WorksAt(Sub, Obj)		p
Jeff	Stanford	0.6
	Princeton	0.4

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

▶ Query Semantics: (“Marginal Probabilities”)

- Run query Q against each instance \mathbf{D}_i ; for each answer tuple t , sum up the probabilities of all instances \mathbf{D}_i where t is a result.

Probabilistic & Temporal Databases

A temporal-probabilistic database \mathbf{D}^{TP} (compactly) encodes a probability distribution over a finite set of deterministic database instances \mathbf{D}_i at each time point of a finite time domain Ω^T .

BornIn(Sub,Obj)		T	p
DeNiro	Green-which	[1943, 1944)	0.9
DeNiro	Tribeca	[1998, 1999)	0.6

Wedding(Sub,Obj)		T	p
DeNiro	Abbott	[1936, 1940)	0.3
DeNiro	Abbott	[1976, 1977)	0.7

Divorce(Sub,Obj)		T	p
DeNiro	Abbott	[1988, 1989)	0.8

- ▶ Sequenced Semantics & Snapshot Reducibility: [Dignös, Gamper, Böhnen: SIGMOD'12]
 - ▶ 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.

$$\text{marriedTo}(x,y)[t_{b1}, T_{\max}) \Leftarrow \text{wedding}(x,y)[t_{b1}, t_{e1}) \wedge \neg \text{divorce}(x,y)[t_{b2}, t_{e2})$$

Sequenced Semantics: Example

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

	R		
	N	D	T
r_1	Joe	DB	[Feb, Jul)
r_2	Ann	DB	[Feb, Sep)
r_3	Sam	AI	[May, Oct)

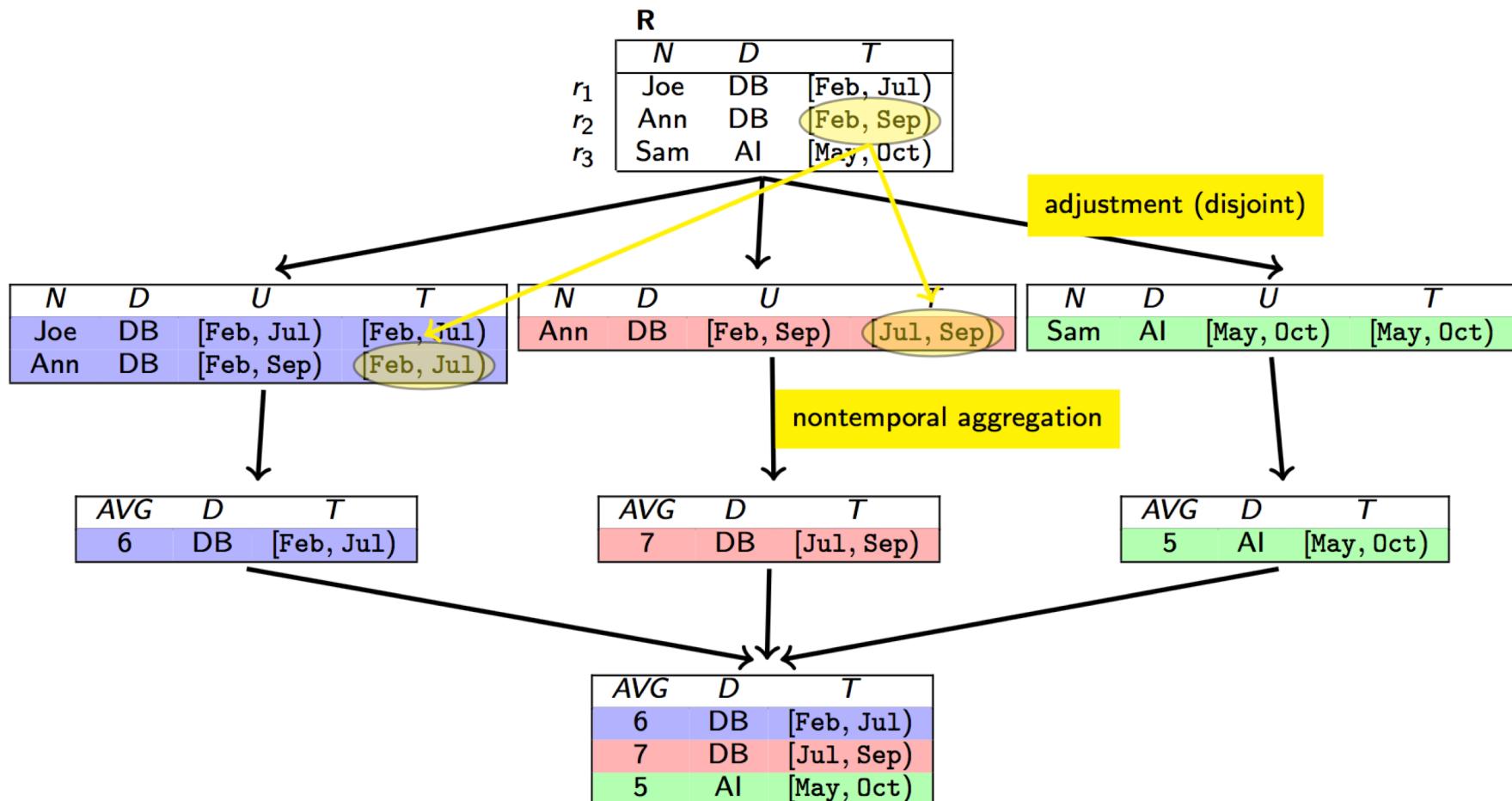
- ▶ Query: How did the average duration of contracts per department change?
- ▶ Result: Temporal Aggregation: $D^{\vartheta^T_{AVG(DUR(T))}}(R)$

	AVG	D	T
z_1	6	DB	[Feb, Jul)
z_2	7	DB	[Jul, Sep)
z_3	5	AI	[May, Oct)

Timestamps must be adjusted for the result.

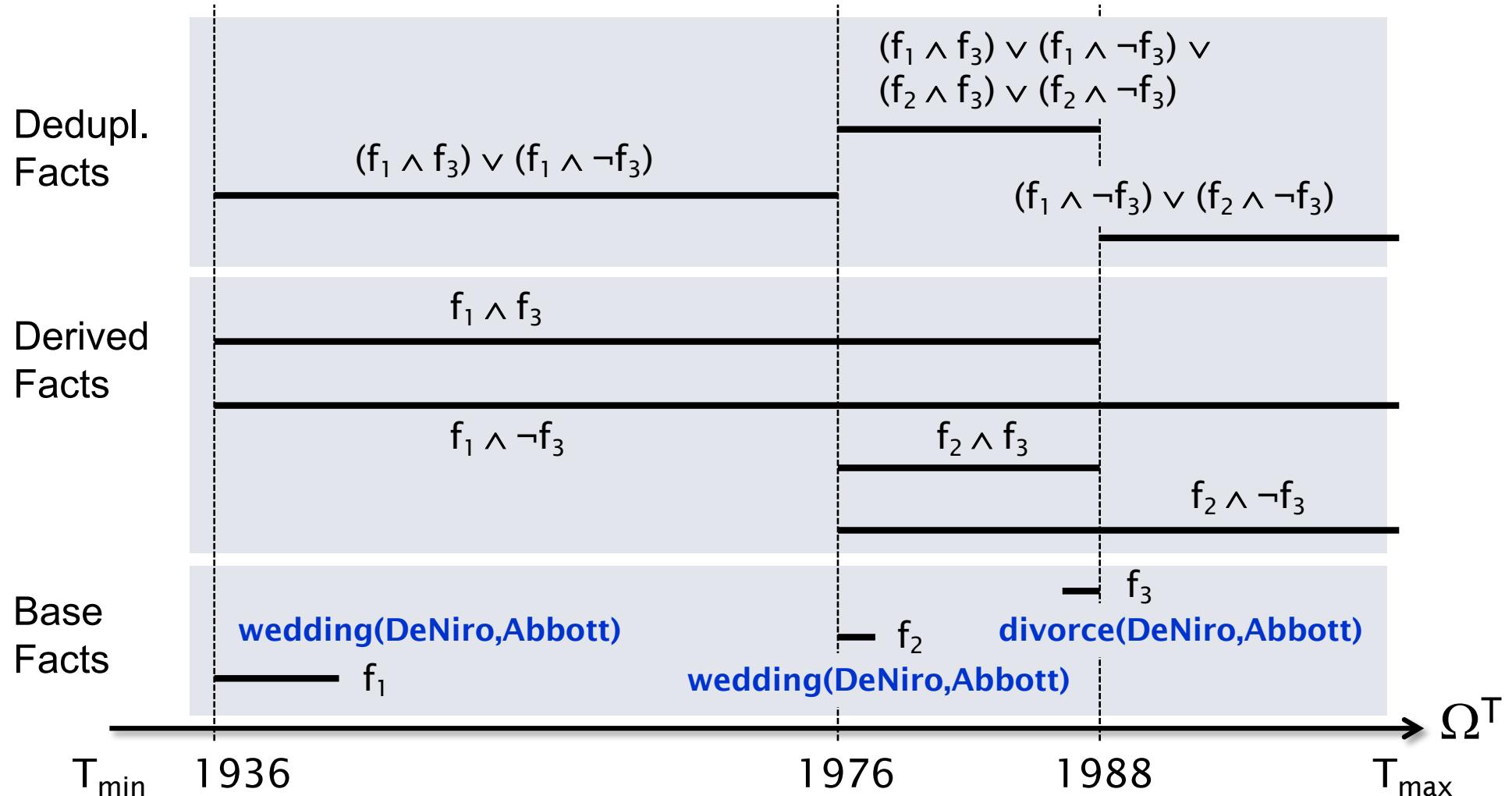
Temporal Splitter / Snapshot Reduction

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



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

Temporal Alignment & Deduplication Example



Non-Sequenced Semantics:

$\text{marriedTo}(x,y)[t_{b1}, T_{\max}] \Leftarrow \text{wedding}(x,y)[t_{b1}, t_{e1}] \wedge \neg \text{divorce}(x,y)[t_{b2}, t_{e2}]$

$\text{marriedTo}(x,y)[t_{b1}, t_{e2}] \Leftarrow \text{wedding}(x,y)[t_{b1}, t_{e1}] \wedge \text{divorce}(x,y)[t_{b2}, t_{e2}] \wedge t_{e1} \leq^T t_{b2}$

Inference in Probabilistic-Temporal Databases

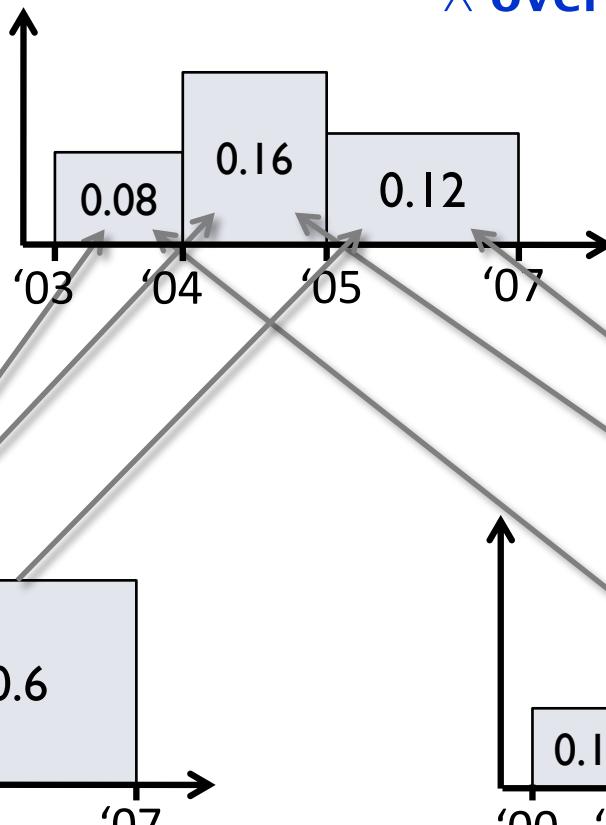
[Wang,Yahya/Theobald: MUD'10; Dylla,Miliaraki/Theobald: PVLDB'13]

Derived
Facts

teamMates(Beckham,
Ronaldo, T_3)



playsFor(Beckham, Real, T_1)
 \wedge playsFor(Ronaldo, Real, T_2)
 \wedge overlaps(T_1 , T_2 , T_3)



Base
Facts

playsFor(Beckham, Real, T_1)

playsFor(Ronaldo, Real, T_2)

Example using the Allen predicate *overlaps*

Inference in Probabilistic-Temporal Databases

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

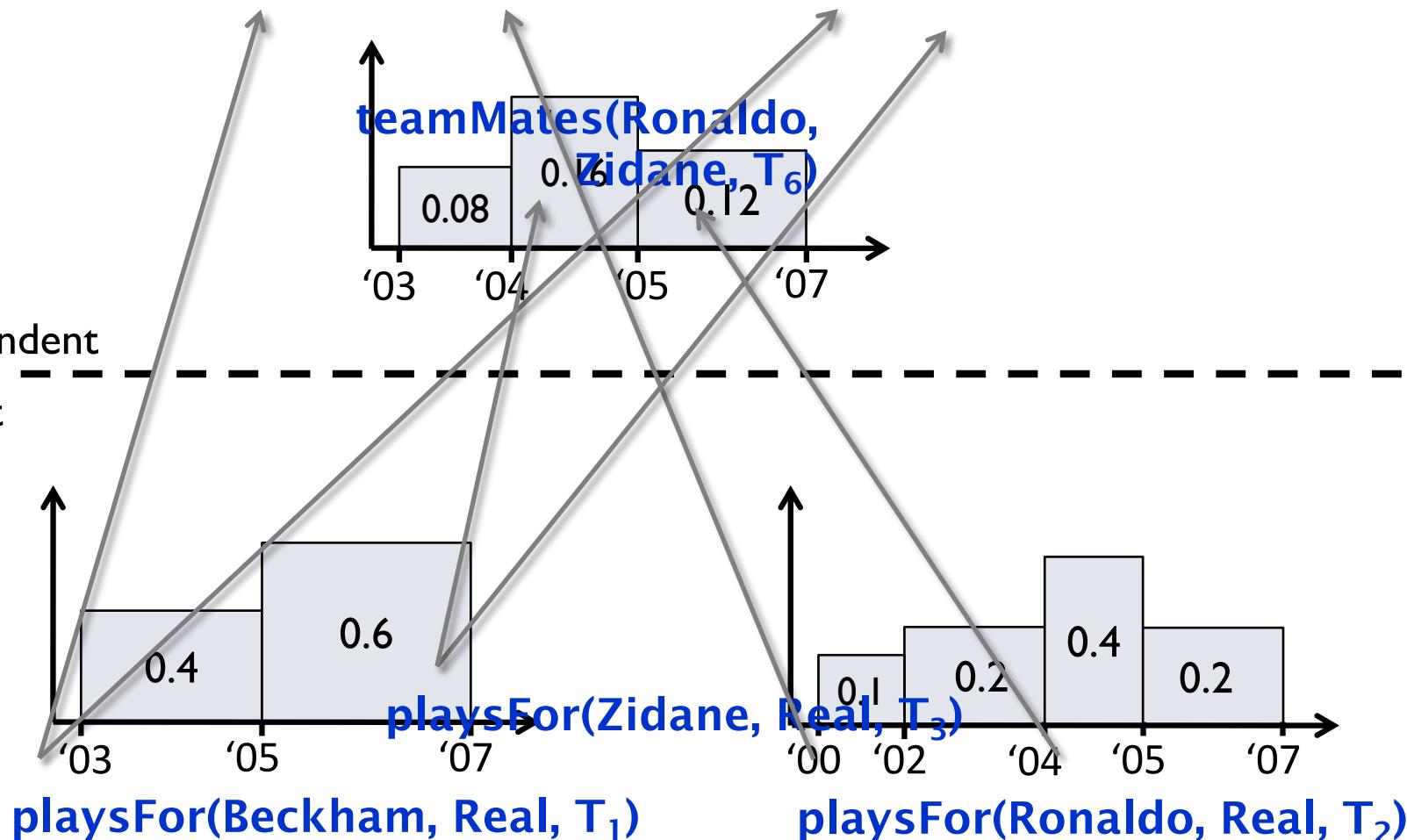
Derived
Facts

teamMates(Beckham,
Ronaldo, T_4)

teamMates(Beckham,
Zidane, T_5)

Non-independent
Independent

Base
Facts



Inference in Probabilistic-Temporal Databases

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

Derived
facts stored
in views

teamMates(Beckham,
Ronaldo, T₄)

teamMates(Beckham,
Zidane, T₅)

teamMates(Ronaldo,
Zidane, T₆)

Non-independent

Independent

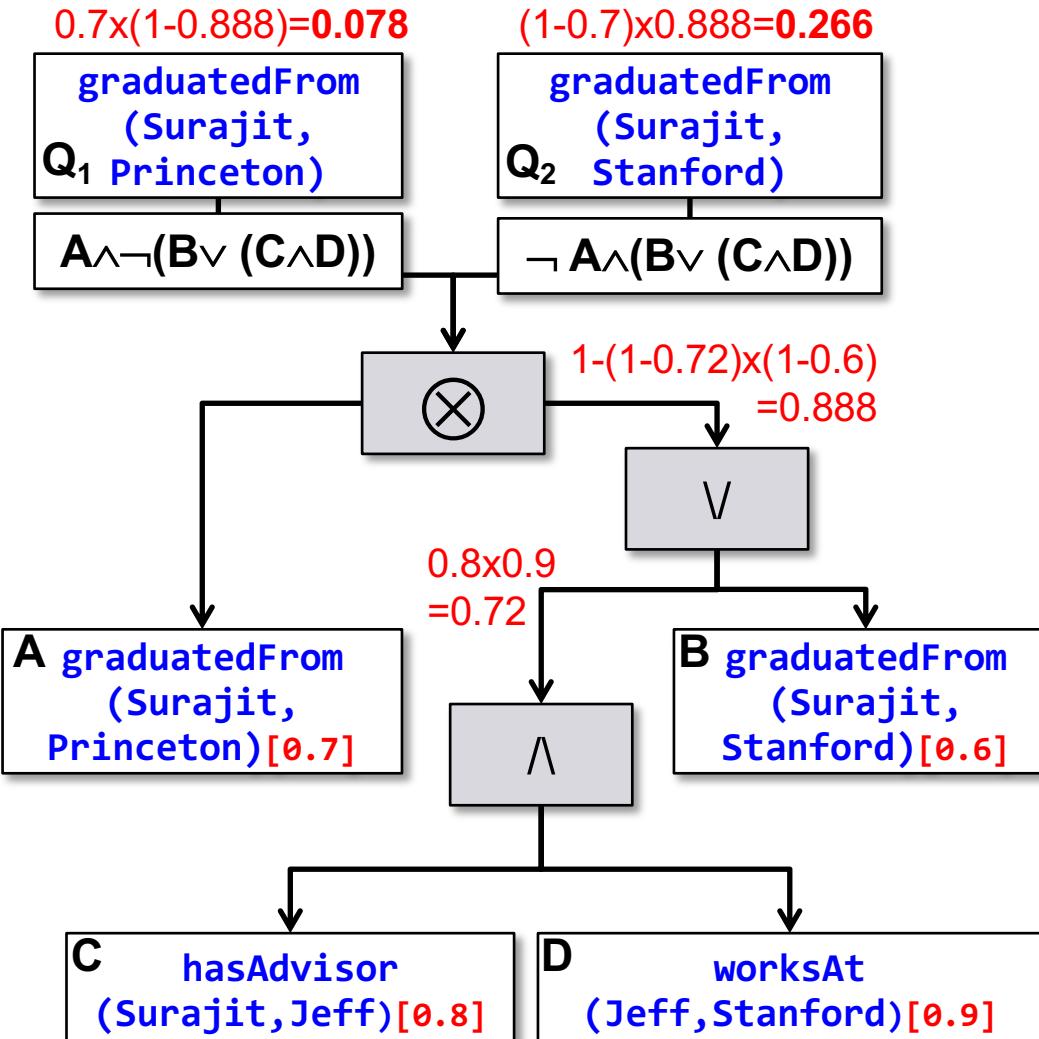


- ▶ **Closed and complete** representation model (incl. lineage)
- ▶ **Temporal alignment** is **polyn.** in the number of input intervals
- ▶ **Confidence computation** per interval remains **#P-hard**
- ▶ In general requires Monte Carlo approximations (Luby-Karp for DNF, MCMC-style sampling), decompositions, or top-k pruning

Lineage & Possible Worlds

Query

`graduatedFrom(Surajit, y)`



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

I) 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”

Literature

[Das Sarma, Theobald, Widom: ICDE'08]

Das Sarma, Anish and Theobald, Martin and Widom, Jennifer, Exploiting Lineage for Confidence Computation in Uncertain and Probabilistic Databases. In: 24th International Conference on Data Engineering (ICDE 2008), IEEE Computer Society Press, pp. 1023-1032. **2008**

[Wang, Yahya, Theobald: MUD'10]

Wang, Yafang and Yahya, Mohamed and Theobald, Martin, Time-aware Reasoning in Uncertain Knowledge Bases. In: 4th International VLDB Workshop on Management of Uncertain Data, Vol. WVP 10-, pp. 51-65, **2010**

[Dignös, Gamper, Böhlen: SIGMOD'12]

A. Dignös, M.H. Böhlen, J. Gamper. Temporal alignment. In Proc. of the SIGMOD-12, pages 433-444, Scottsdale, AZ, USA, May 20-24, **2012**

[Dylla, Miliaraki, Theobald: PVLDB'13]

Maximilian Dylla, Iris Miliaraki, Martin Theobald, A Temporal-Probabilistic Database Model for Information Extraction, Proceedings of the VLDB Endowment, Volume 6, Issue 14, **2013**

Historical Facts vs. Future Facts

- ▶ Processing uncertain historical data

$\text{marriedTo}(x,y)[t_{b1}, t_{e2}] \Leftarrow \text{wedding}(x,y)[t_{b1}, t_{e1}) \wedge \neg \text{divorce}(x,y)[t_{b2}, t_{e2})$

$\text{marriedTo}(x,y)[t_{b1}, t_{e2}) \Leftarrow \text{wedding}(x,y)[t_{b1}, t_{e1}) \wedge \text{divorce}(x,y)[t_{b2}, t_{e2}) \wedge t_{e1} \leq^T t_{b2}$

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

$\text{marriedTo}(x,y)[t_{b1}, T_{\max}] \Leftarrow \text{wedding}(x,y)[t_{b1}, t_{e1}) \wedge \neg \text{divorce}(x,y)[t_{b2}, t_{e2})$

$\text{marriedTo}(x,y)[t_{b1}, t_{e2}) \Leftarrow \text{wedding}(x,y)[t_{b1}, t_{e1}) \wedge \text{divorce}(x,y)[t_{b2}, t_{e2}) \wedge t_{e1} \leq^T t_{b2}$