Constrained Open-World Probabilistic Databases

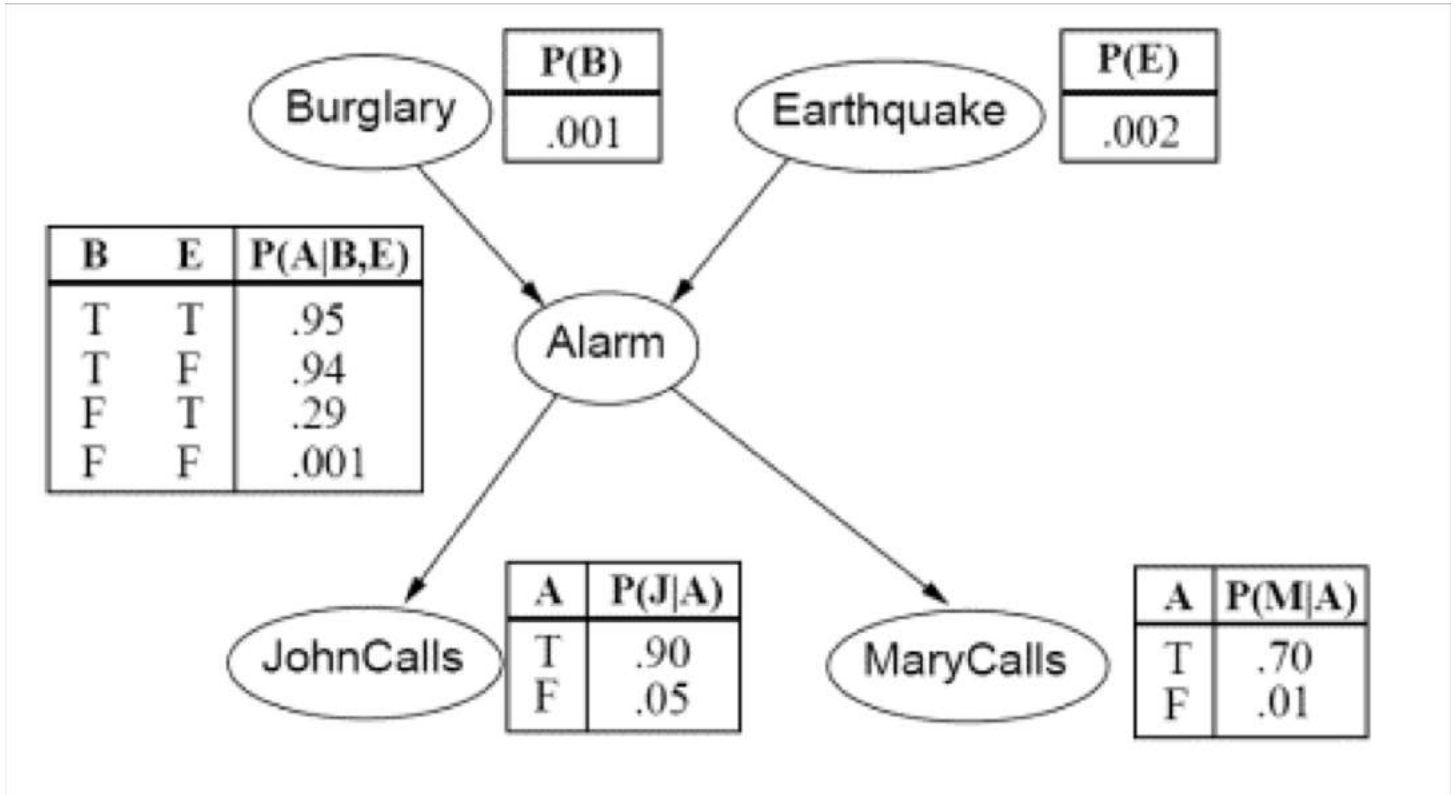
Tal Friedman and Guy Van den Broeck

Goal

- Model relational data with uncertainty
- Ask model interesting questions

What's the challenge?

- Probabilities make things less tractable



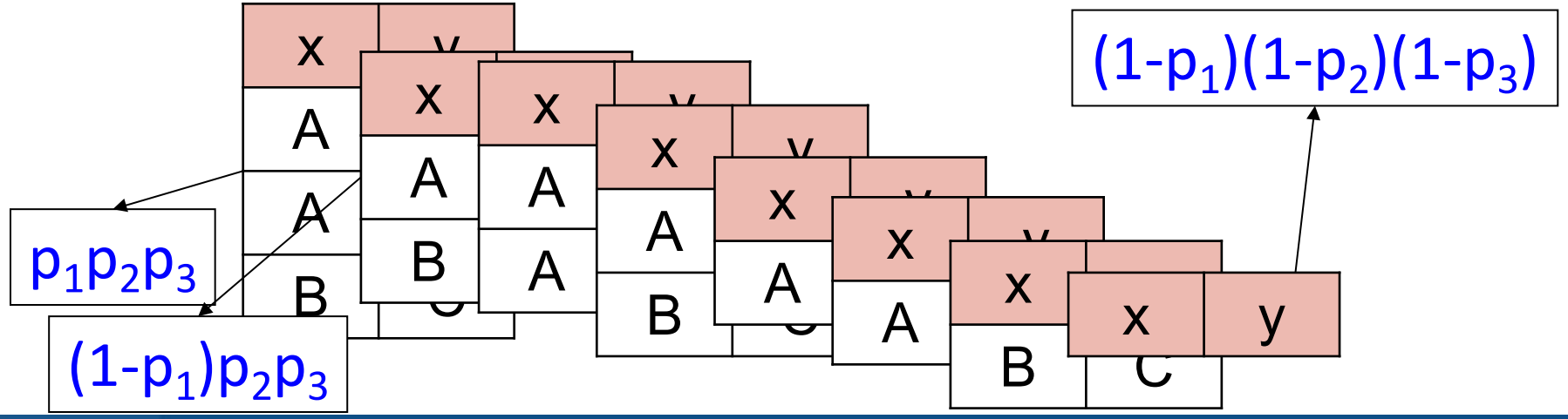
Probabilistic Databases

Probabilistic database D:

Coauthor

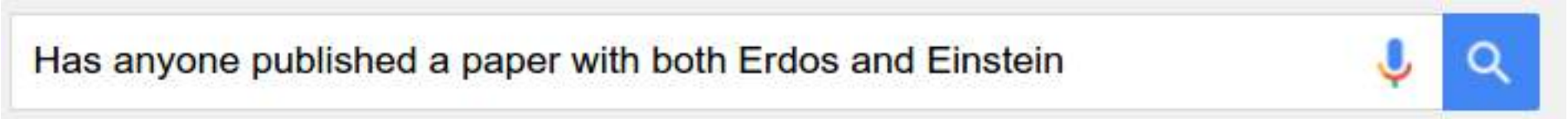
x	y	P
A	B	p_1
A	C	p_2
B	C	p_3

Possible worlds semantics:



[VdB&Suciu'17]

Running Example



- Reason using facts about scientists + coauthorship

Scientist

x	P
Erdos	0.9
Einstein	0.8
Pauli	0.6

Coauthor

x	y	P
Erdos	Renyi	0.6
Einstein	Pauli	0.7
Obama	Erdos	0.1

- Scraped/learned from web, large text corpora

[VdB&Suciu'17]

What’s the challenge?

- Probabilities make life difficult
- Our knowledge is not complete

Open-World Probabilistic Databases

Unknown tuples can be added with probability P <= λ

X	Y	P
Einstein	Straus	0.7
~~**Erdos**~~	~~**Straus**~~	~~**0.6**~~
Einstein	Pauli	0.9
Erdos	Renyi	0.7
Kersting	Natarajan	0.8
Luc	Paol	0.1
…	…	…
Erdos	**Straus**	λ

Open-World Probabilistic Databases

Open-World makes *everything* possible

Want something more meaningful

X	Y	P
Einstein	Straus	0.7
~~**Erdos**~~	~~**Straus**~~	~~**0.6**~~
Einstein	Pauli	0.9
Erdos	Renyi	0.7
Kersting	Natarajan	0.8
Luc	Paol	0.1
…	…	…
Erdos	**Straus**	**λ**
Bieber	**Einstein**	**λ**
Friedman	**Bieber**	**λ**
Banner	**Friedman**	**λ**
…	…	…

Open-World Probabilistic Databases

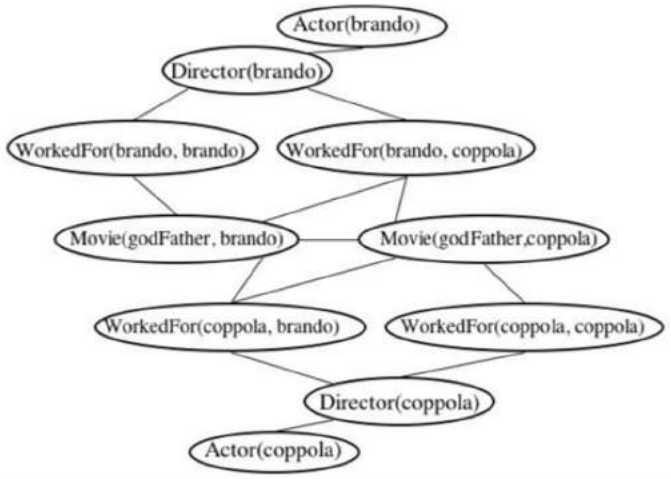
~~Open-World makes *everything* possible~~

Constrain to “reasonable” options

(a)

0.7	$Actor(A) \Rightarrow \neg Director(A)$
1.2	$Director(A) \Rightarrow \neg WorkedFor(A, B)$
1.4	$Movie(T, A) \wedge WorkedFor(A, B) \Rightarrow Movie(T, B)$

(b)



Not tractable

Hard to construct

X	Y	P
Einstein	Straus	0.7
~~**Erdos**~~	~~**Straus**~~	~~**0.6**~~
Einstein	Pauli	0.9
Erdos	Renyi	0.7
Kersting	Natarajan	0.8
Luc	Paol	0.1
…	…	…
Erdos	**Straus**	**λ**
Bieber	**Einstein**	**λ**
Friedman	**Bieber**	**λ**
Banner	**Friedman**	**λ**
…	…	…

Constrained Open-World Probabilistic Databases

~~Open-World makes *everything* possible~~

Constrain to “reasonable” options

Just use a summary statistic!

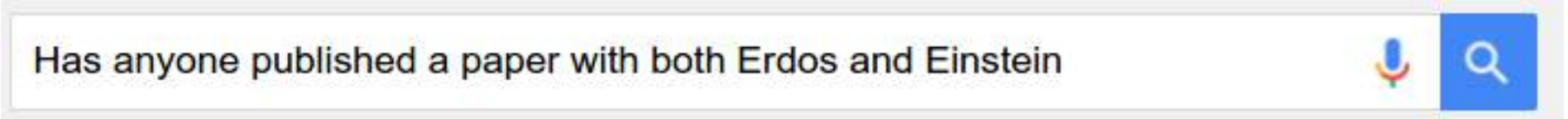
X	Y	P
Einstein	Straus	0.7
~~**Erdos**~~	~~**Straus**~~	~~**0.6**~~
Einstein	Pauli	0.9
Erdos	Renyi	0.7
Kersting	Natarajan	0.8
Luc	Paol	0.1
…	…	…
Erdos	**Straus**	**λ**
Bieber	**Einstein**	**λ**
Friedman	**Bieber**	**λ**
Banner	**Friedman**	**λ**
…	…	…

$$\frac{1}{n}\sum \leq p$$

Constrained Open-World Probabilistic Databases

1. Identify a class of tractable queries with algorithm
2. Outline where querying becomes more difficult
3. Provide an efficient approximation

Running Example



- Reason using facts about scientists + coauthorship

Scientist

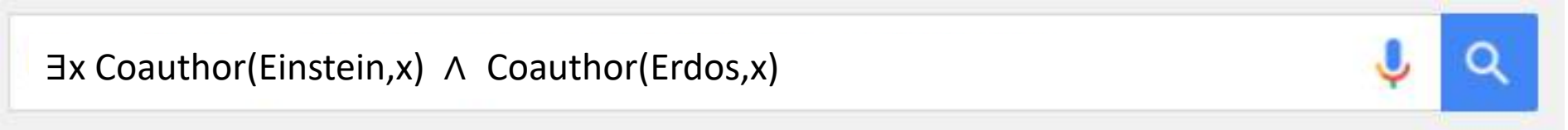
x	P
Erdos	0.9
Einstein	0.8
Pauli	0.6

Coauthor

x	y	P
Erdos	Renyi	0.6
Einstein	Pauli	0.7
Obama	Erdos	0.1

[VdB&Suciu'17]

Queries



- Conjunctive queries (CQ): ∃ + ∧ + positive literals
- Unions of conjunctive queries (UCQ): v of CQs

Query Evaluation

- Computing query probability forms a *dichotomy*:
  - PTIME (linear), *safe* queries
  - #P-hard *unsafe* queries

- Can be symbolically determined!

Constrained Open-World Probabilistic Databases

Querying is now an optimization problem

Constrained Open-World Probabilistic Databases

Querying is now an optimization problem

Select p_i ‘s such that:

- $\sum p_i \leq p$
- $0 \leq p_i \leq \lambda \forall i$
- Query probability is maximized

Constrained Open-World Probabilistic Databases

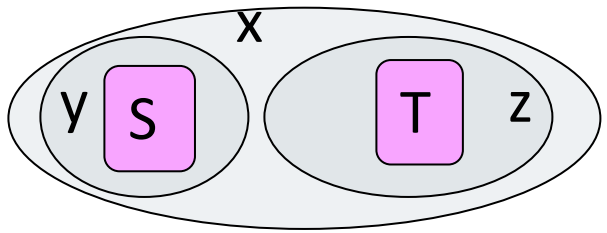
1. **Identify a class of tractable queries with algorithm**
2. Outline where querying becomes more difficult
3. Provide an efficient approximation

Tractable Queries

Tractability typically depends on a hierarchical property

Hierarchical

$Q = \exists x \exists y \exists z\ S(x,y) \wedge T(x,z)$



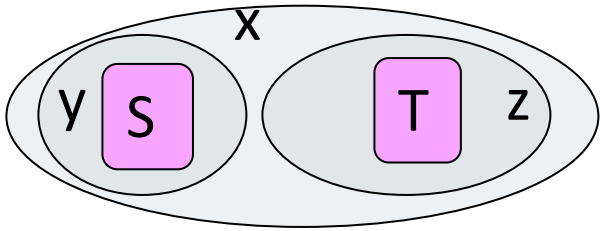
[VdB'18]

Tractable Queries

Tractability depends on a hierarchical property

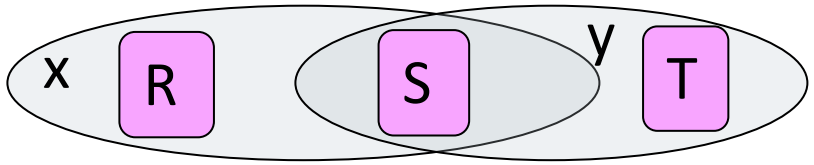
Hierarchical

$Q = \exists x \exists y \exists z\ S(x,y) \wedge T(x,z)$



Non-hierarchical

$H_0 = \exists x \exists y\ (R(x) \wedge S(x,y) \wedge T(y))$



[VdB'18]

Tractable Queries

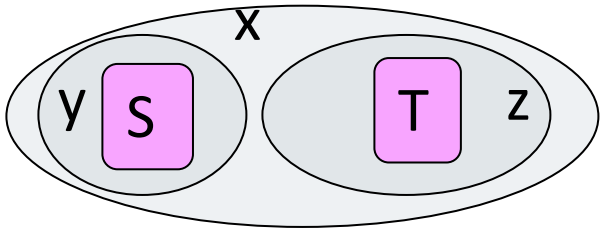
Tractability depends on a hierarchical property

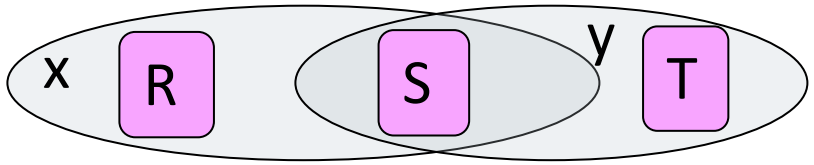
Hierarchical

$Q = \exists x \exists y \exists z\ S(x,y) \wedge T(x,z)$

Non-hierarchical

$H_0 = \forall x \forall y\ (R(x) \wedge S(x,y) \wedge T(y))$





If all CQs in a UCQ are hierarchical, the query is safe

[VdB'18]

Tractable Queries

If all CQs in a UCQ are hierarchical, the query is safe

Tractable Queries

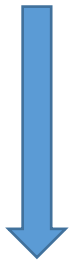
~~If all CQs in a UCQ are hierarchical, the query is safe~~

With constraints, all CQs need to have the same hierarchy

Tractable Queries

~~If all CQs in a UCQ are hierarchical, the query is safe~~

With constraints, all CQs need to have the same hierarchy



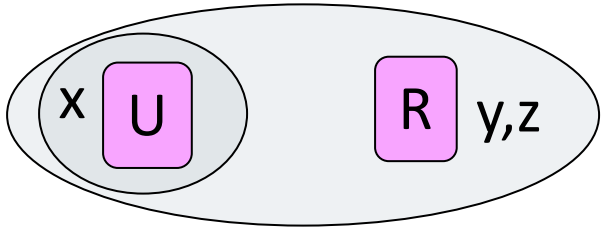
Efficient dynamic programming algorithm

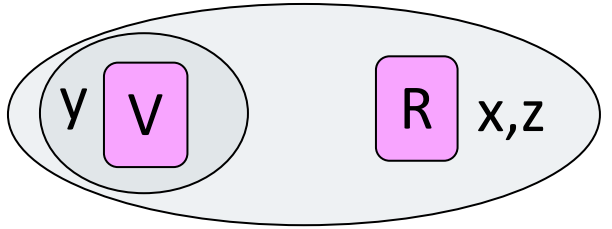
Constrained Open-World Probabilistic Databases

1. Identify a class of tractable queries with algorithm
2. **Outline where querying becomes more difficult**
3. Provide an efficient approximation

Constraints change the hardness landscape

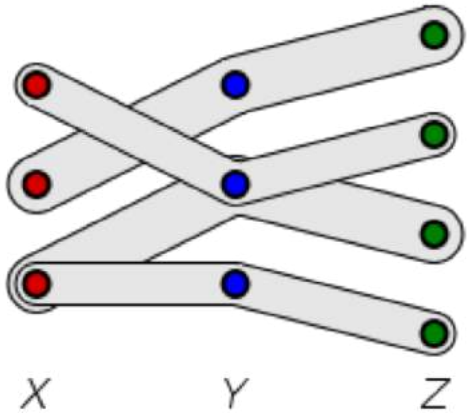
$$M_0 = \exists x \exists y \exists z \, (R(x,y,z) \wedge U(x)) \vee (R(x,y,z) \wedge V(y)) \\ \vee (R(x,y,z) \wedge W(z)) \vee (U(x) \wedge V(y)) \\ \vee (U(x) \wedge W(z)) \vee (V(y) \wedge W(z))$$





Query is *safe* for PDB/OpenPDB evaluation

Constraints make it NP-hard



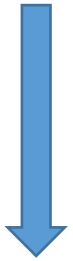
Constrained Open-World Probabilistic Databases

1. Identify a class of tractable queries with algorithm
2. Outline where querying becomes more difficult
3. **Provide an efficient approximation**

Approximation

Consider $f(S)$: query prob. if we give all tuples in S prob. λ

f is monotonic and submodular



Efficient + accurate greedy approximation!

X	Y	P
Einstein	Straus	0.7
~~**Erdos**~~	~~**Straus**~~	~~**0.6**~~
Einstein	Pauli	0.9
Erdos	Renyi	0.7
Kersting	Natarajan	0.8
Luc	Paol	0.1
…	…	…
Erdos	**Straus**	λ

$f\left(\text{CoA(Erdos, Straus)}\right)$

Conclusion

- Modelling uncertainty when managing large amounts of data requires unreasonably strong assumptions
- We show how to make these models more realistic without any additional row level information