

Open-World Probabilistic Databases

Guy Van den Broeck

UCLA



FLAIRS
May 23, 2017

Overview

1. *Why probabilistic databases?*
2. *How probabilistic query evaluation?*
3. *Why open world?*
4. *How open-world query evaluation?*
5. *What is the broader picture?*

Why probabilistic databases?

What we'd like to do...

Has anyone published a paper with both Erdos and Einstein



All

News

Images

Videos

Shopping

More ▾

Search tools

About 82,400 results (0.73 seconds)

Erdős number - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Erdős_number ▾ Wikipedia ▾

He **published** more **papers** during his lifetime (at least 1,525) than any other ...

Anybody else's Erdős number is $k + 1$ where k is the lowest Erdős number of any coauthor. ... **Albert Einstein and Sheldon Lee Glashow have an Erdős number of 2.** ...

and mathematician Ruth Williams, **both** of whom **have** an Erdős number of 2.

Erdős–Bacon number - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Erdős–Bacon_number ▾ Wikipedia ▾

This article possibly **contains** previously unpublished synthesis of **published** ... Her **paper** gives her an Erdős number of 4, and a Bacon number of 2, **both** of ...

Google Knowledge Graph

The image shows a Google search interface for 'Larry Page'. The search bar contains the text 'Larry Page'. Below the search bar, there are tabs for 'Web', 'Images', 'Maps', 'Shopping', 'News', and 'More -'. The search results show a snippet about 'Ubergizmo' and several links to 'Larry Page - Forbes', 'Larry Page - Google+', 'Management team - Company - Google', 'Larry Page Biography - Facts, Birthday, Life Story - Biography.com', and 'Larry Page | CrunchBase Profile'. On the right side, there is a 'Knowledge Graph' panel for Larry Page, which includes a large portrait, a grid of smaller images, and a list of biographical facts such as birth date, height, spouse, siblings, education, and awards. The panel also shows 'Recent posts' and 'People also search for'.

Knowledge Graph

> 570 million entities
> 18 billion tuples

Ubergizmo - 3 days ago
Android 4.4 KitKat marks a milestone for Google as they have named their mobile operating system after a branded chocolate – although ...

[Larry Page - Forbes](#)
www.forbes.com/profile/larry-page/
Larry Page on Forbes - #20 Billionaires, #20 Powerful People, #13 Forbes 400.

[Larry Page - Google+](#)
<https://plus.google.com/+LarryPage>
by Larry Page - in 6,606,272 Google+ circles
Dear Google users— You may be aware of press reports alleging that Internet companies have joined a secret U.S. government program called PRISM to give ...

[Management team - Company - Google](#)
www.google.com/about/company/facts/management/
Larry Page and Sergey Brin founded Google in September 1998. Since then, the company has grown to more than 30,000 employees worldwide, with a ...

[Larry Page Biography - Facts, Birthday, Life Story - Biography.com](#)
www.biography.com > People
You don't need a search engine to find out all there is to know about Larry Page, co-founder of Google. Just come to Biography.com!

[Larry Page | CrunchBase Profile](#)
www.crunchbase.com > People
Larry Page was Google's founding CEO and grew the company to more than 200 employees and profitability before moving into.

Larry Page
6,606,633 followers on Google+

Lawrence "Larry" Page is an American computer scientist and Internet entrepreneur who is the co-founder of Google, alongside Sergey Brin. On April 4, 2011, Page succeeded Eric Schmidt as the chief executive officer of Google. [Wikipedia](#)

Born: March 26, 1973 (age 40), East Lansing, MI
Height: 5' 11" (1.80 m)
Spouse: Lucinda Southworth (m. 2007)
Siblings: Carl Victor Page, Jr.
Education: East Lansing High School (1987–1991), More
Awards: Marconi Prize, TR100

Recent posts
Just opened the new Android release. KitKat! Sep 3, 2013

People also search for

Probabilistic Databases

Has anyone published a paper with both Erdos and Einstein



- Tuple-independent probabilistic database

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

- Learned from the web, large text corpora, ontologies, etc., using **statistical** machine learning.

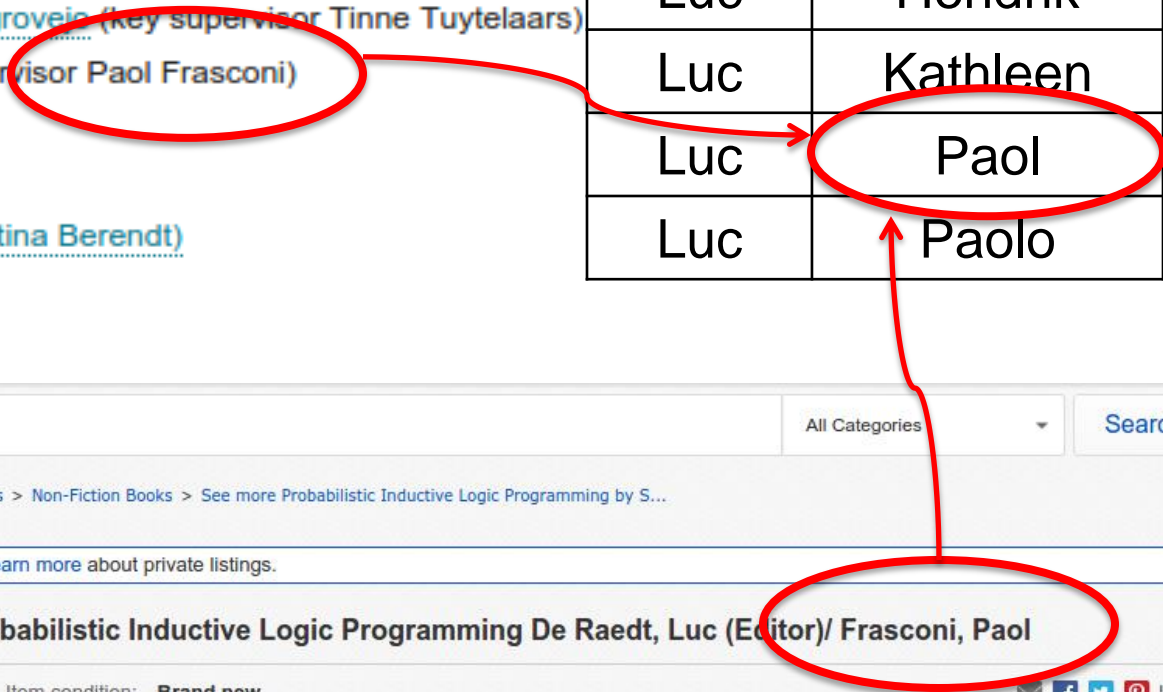
Information Extraction is Noisy!

PhD Students Luc De Raedt

- [Laura-Andrea Antanas](#) (co-promotor Tinne Tuytelaars)
- [Dries Van Daele](#) (co-promotor Kathleen Marchal)
- [Thanh Le Van](#) (co-promotor Kathleen Marchal)
- [Bogdan Moldovan](#)
- [Davide Nitti](#) (co-promotor Tinne De Laet)
- [José Antonio Oramas Mogrovejo](#) (key supervisor Tinne Tuytelaars)
- [Francesco Orsini](#) (co-supervisor **Paol Frasconi**)
- [Sergey Paramonov](#)
- [Joris Renkens](#)
- [Mathias Verbeke](#) (with Bettina Berendt)
- [Jonas Vlasselaer](#)

Coauthor

x	y	P
Luc	Laura	0.7
Luc	Hendrik	0.6
Luc	Kathleen	0.3
Luc	Paol	0.3
Luc	Paolo	0.1



ebay Shop by category Search... All Categories Search Advanced

Back to home page | Listed in category: Books, Magazines > Non-Fiction Books > See more Probabilistic Inductive Logic Programming by S...

This is a private listing. Sign in to view your status or learn more about private listings.

Probabilistic Inductive Logic Programming De Raedt, Luc (Editor)/ Frasconi, Paol

Item condition: **Brand new** Time left: 18d 13h (22 Feb, 2016 04:40:52 AEDST)

Seller information

What we'd like to do...

$\exists x \text{ Coauthor}(\text{Einstein}, x) \wedge \text{ Coauthor}(\text{Erdos}, x)$



All

News

Images

Videos

Shopping

More ▾

Search tools

About 82,400 results (0.73 seconds)

Erdős number - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Erdős_number ▾ Wikipedia ▾

He **published** more **papers** during his lifetime (at least 1,525) than any other ...

Anybody else's **Erdős** number is $k + 1$ where k is the lowest **Erdős** number of any coauthor. ... Albert **Einstein** and Sheldon Lee Glashow **have** an **Erdős** number of 2. ... and mathematician Ruth Williams, **both** of whom **have** an **Erdős** number of 2.

Erdős–Bacon number - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Erdős–Bacon_number ▾ Wikipedia ▾

This article possibly **contains** previously unpublished synthesis of **published** ... Her **paper** gives her an **Erdős** number of 4, and a Bacon number of 2, **both** of ...

Einstein is in the Knowledge Graph

Albert Einstein



All News Images Books Videos More Search tools

About 82,800,000 results (0.45 seconds)

The Official Licensing Site of Albert Einstein

einstein.biz/

Welcome to the Official Licensing Site of **Albert Einstein**. Learn more about **Albert Einstein** and contact us today for any commercial licensing inquiries.

Albert Einstein - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Albert_Einstein

Albert Einstein (/ˈaɪnʃtaɪn/; German: [ˈalbɛʁt ˈaɪnʃtaɪn] (listen); 14 March 1879 – 18 April 1955) was a German-born theoretical physicist.

[Hans Albert Einstein](#) - [Mass–energy equivalence](#) - [Eduard Einstein](#) - [Elsa Einstein](#)

Albert Einstein (@AlbertEinstein) | Twitter

<https://twitter.com/AlbertEinstein>

16 hours ago - [View on Twitter](#)

ICYMI, Albert Einstein knew a thing or two about being romantic. Learn about the love letters he wrote. guff.com/didnt-know-einst...

20 hours ago - [View on Twitter](#)

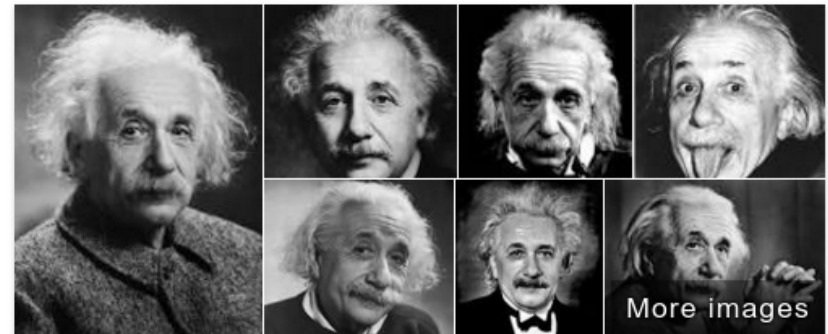
An interesting read on Einstein's superstar status. What are your thoughts? twitter.com/aeonmag/status...



Albert Einstein - Biographical - Nobelprize.org

www.nobelprize.org/nobel_prizes/physics/.../einstein-bio.htm...

Albert Einstein was born at Ulm, in Württemberg, Germany, on March 14, 1879. ... Later, they moved to Italy and Albert continued his education at Aarau



Albert Einstein

Theoretical Physicist

Albert Einstein was a German-born theoretical physicist. He developed the general theory of relativity, one of the two pillars of modern physics. Einstein's work is also known for its influence on the philosophy of science. [Wikipedia](#)

Born: March 14, 1879, [Ulm, Germany](#)

Died: April 18, 1955, [Princeton, NJ](#)

Influenced by: [Isaac Newton](#), [Mahatma Gandhi](#), [More](#)

Children: [Eduard Einstein](#), [Lieserl Einstein](#), [Hans Albert Einstein](#)

Spouse: [Elsa Einstein](#) (m. 1919–1936), [Mileva Marić](#) (m. 1903–1919)

Erdős is in the Knowledge Graph

Paul Erdos



All Images Videos Books News More Search tools

About 333,000 results (0.35 seconds)

Paul Erdős - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Paul_Erdős - Wikipedia

Paul Erdős was a Hungarian Jewish mathematician. He was one of the most prolific mathematicians of the 20th century. He was known both for his social ...

Fan Chung - Ronald Graham - Béla Bollobás - Category:Paul Erdős

The Man Who Loved Only Numbers - The New York Times

<https://www.nytimes.com/books/.../hoffman-man.ht...> - The New York Times

Paul Erdős was one of those very special geniuses, the kind who comes along only once in a very long while yet he chose, quite consciously I am sure, to share ...

Paul Erdos | Hungarian mathematician | Britannica.com

www.britannica.com/biography/Paul-Erdos - Encyclopaedia Britannica

Paul Erdős, (born March 26, 1913, Budapest, Hungary—died September 20, 1996, Warsaw, Poland), Hungarian “freelance” mathematician (known for his work ...

Paul Erdős - University of St Andrews

www-groups.dcs.st-and.ac.uk/~history/Biographies/Erdos.html

Paul Erdős came from a Jewish family (the original family name being Engländer) although neither of his parents observed the Jewish religion. Paul's father ...

[PDF] Paul Erdős Mathematical Genius, Human - UnTruth.org

www.untruth.org/~josh/math/Paul%20Erdős%20bio-rev2.pdf

by J Hill - 2004 - Related articles



Paul Erdős

Mathematician

Paul Erdős was a Hungarian Jewish mathematician. He was one of the most prolific mathematicians of the 20th century. He was known both for his social practice of mathematics and for his eccentric lifestyle.

[Wikipedia](#)

Born: March 26, 1913, Budapest, Hungary

Died: September 20, 1996, Warsaw, Poland

Education: Eötvös Loránd University (1934)

Books: Probabilistic Methods in Combinatorics, More

Notable students: Béla Bollobás, Alexander Soifer, George B. Purdy, Joseph Kruskal

This guy is in the Knowledge Graph

Ernst Straus



All Images News Maps Videos More Search tools

About 349,000 results (0.37 seconds)

Ernst G. Straus - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Ernst_G._Straus Wikipedia

Ernst Gabor Straus (February 25, 1922 – July 12, 1983) was a German-American mathematician who helped found the theories of Euclidean Ramsey theory ...

Straus biography - University of St Andrews

www-groups.dcs.st-and.ac.uk/~history/Biographies/Straus.html

Ernst Straus's mother was Rahel Goitein who had the distinction of being one of the first women medical students officially studying at a German university.

Images for Ernst Straus

Ernst G. Straus

Mathematician

Ernst Gabor Straus was a German-American mathematician who helped found the theories of Euclidean Ramsey theory and of the arithmetic properties of analytic functions. [Wikipedia](#)

Born: February 25, 1922, [Munich, Germany](#)

Died: July 12, 1983, [Los Angeles, CA](#)

Residence: [United States of America](#)

... and he published with both Einstein and Erdos!

Desired Query Answer

Has anyone published a paper with both Erdos and Einstein



Ernst Straus



Barack Obama, ...



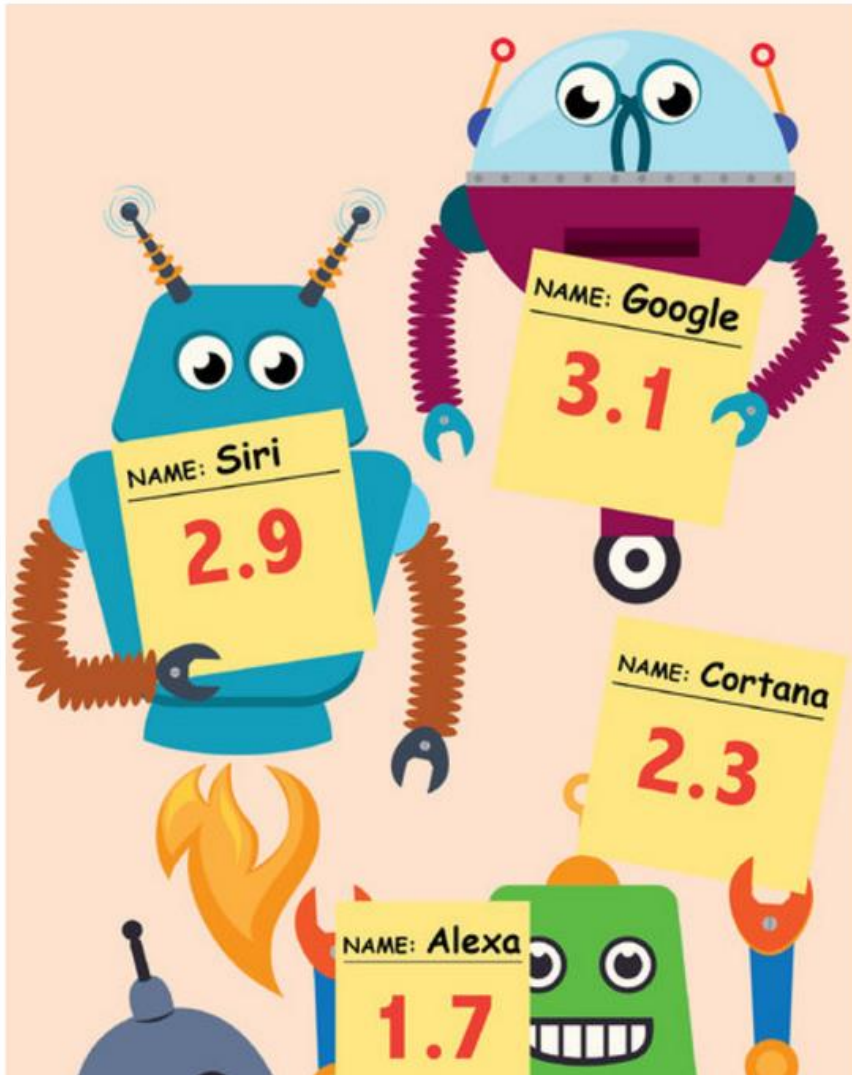
Justin Bieber, ...

1. Fuse uncertain information from web
⇒ **Embrace probability!**
2. Cannot come from labeled data
⇒ **Embrace query eval!**

Siri, Alexa and Other Virtual Assistants Put to the Test

Tech Fix

By BRIAN X. CHEN JAN. 27, 2016



WHEN I asked Alexa earlier this week who was playing in the [Super Bowl](#), she responded, somewhat monotonously, “[Super Bowl](#) 49’s winner is New England Patriots.”

“Come on, that’s last year’s Super Bowl,” I said. “Even I can do better than that.”

At the time, I was actually alone in my living room. I was talking to the virtual companion inside [Amazon](#)’s wireless speaker, Echo, which was released last June. Known as Alexa, she has gained raves from Silicon Valley’s tech-obsessed digerati and has become one of the newest members of the virtual assistants club.

All the so-called [Frightful Five](#) tech

[Chen’16]
(NYTimes)

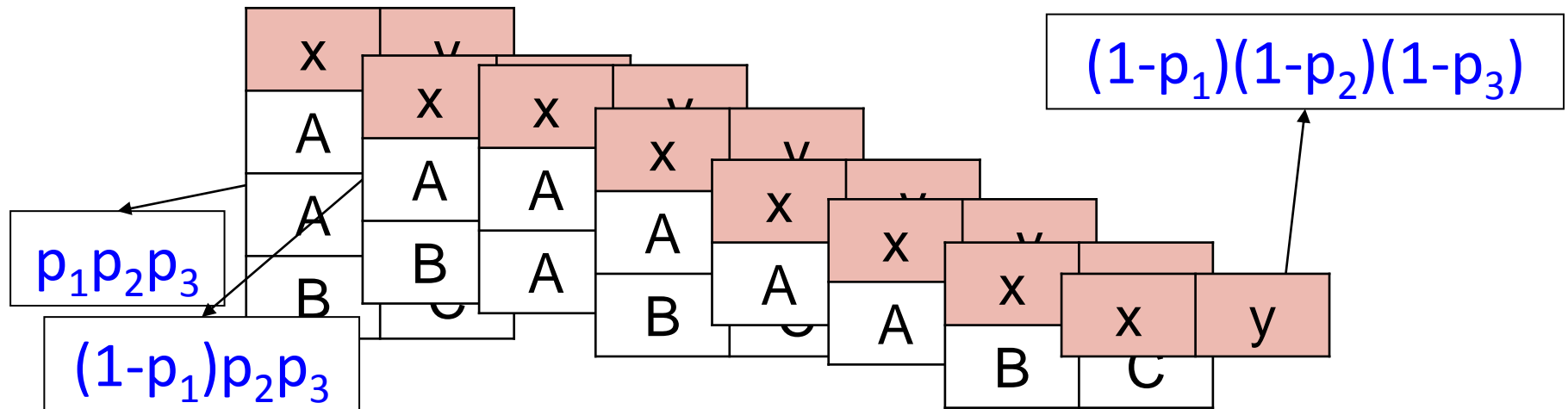
***How probabilistic
query evaluation?***

Tuple-Independent Probabilistic DB

Probabilistic database D:

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

Possible worlds semantics:



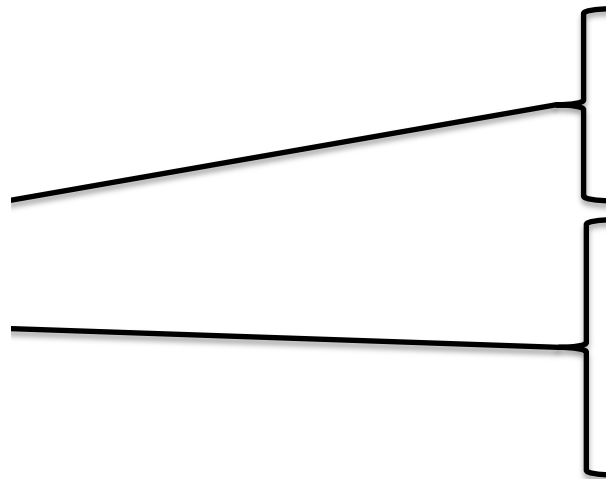
Probabilistic Query Evaluation

$$Q = \exists x \exists y \text{ Scientist}(x) \wedge \text{Coauthor}(x,y)$$

$$P(Q) = 1 - \{1 - p_1 * [1 - (1 - q_1) * (1 - q_2)]\} * \\ \{1 - p_2 * [1 - (1 - q_3) * (1 - q_4) * (1 - q_5)]\}$$

Scientist

x	P
A	p_1
B	p_2
C	p_3



x	y	P
A	D	q_1
A	E	q_2
B	F	q_3
B	G	q_4
B	H	q_5

Coauthor

Lifted Inference Rules

Preprocess Q (omitted),
Then apply rules (some have preconditions)

$$P(\neg Q) = 1 - P(Q)$$

Negation

$$P(Q1 \wedge Q2) = P(Q1) P(Q2)$$
$$P(Q1 \vee Q2) = 1 - (1 - P(Q1)) (1 - P(Q2))$$

Decomposable \wedge, \vee

$$P(\exists z Q) = 1 - \prod_{A \in \text{Domain}} (1 - P(Q[A/z]))$$
$$P(\forall z Q) = \prod_{A \in \text{Domain}} P(Q[A/z])$$

Decomposable \exists, \forall

$$P(Q1 \wedge Q2) = P(Q1) + P(Q2) - P(Q1 \vee Q2)$$
$$P(Q1 \vee Q2) = P(Q1) + P(Q2) - P(Q1 \wedge Q2)$$

Inclusion/
exclusion

Closed-World Lifted Query Eval

$$Q = \exists x \exists y \text{Scientist}(x) \wedge \text{Coauthor}(x,y)$$

Decomposable \forall -Rule

$$P(Q) = 1 - \prod_{A \in \text{Domain}} (1 - P(\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y)))$$

$$\begin{aligned} &= 1 - (1 - P(\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y))) \\ &\quad \times (1 - P(\text{Scientist}(B) \wedge \exists y \text{Coauthor}(B,y))) \\ &\quad \times (1 - P(\text{Scientist}(C) \wedge \exists y \text{Coauthor}(C,y))) \\ &\quad \times (1 - P(\text{Scientist}(D) \wedge \exists y \text{Coauthor}(D,y))) \\ &\quad \times (1 - P(\text{Scientist}(E) \wedge \exists y \text{Coauthor}(E,y))) \\ &\quad \times (1 - P(\text{Scientist}(F) \wedge \exists y \text{Coauthor}(F,y))) \end{aligned}$$

...

Check independence:

$$\begin{aligned} &\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y) \\ &\text{Scientist}(B) \wedge \exists y \text{Coauthor}(B,y) \end{aligned}$$

Complexity PTIME

Limitations

$$H_0 = \forall x \forall y \text{ Smoker}(x) \vee \text{Friend}(x,y) \vee \text{Jogger}(y)$$

The decomposable \forall -rule:
... does not apply:

$$P(\forall z Q) = \prod_{A \in \text{Domain}} P(Q[A/z])$$

$H_0[\text{Alice}/x]$ and $H_0[\text{Bob}/x]$ are dependent:



$$\forall y (\text{Smoker}(\text{Alice}) \vee \text{Friend}(\text{Alice},y) \vee \text{Jogger}(y))$$

$$\forall y (\text{Smoker}(\text{Bob}) \vee \text{Friend}(\text{Bob},y) \vee \text{Jogger}(y))$$

Lifted inference sometimes fails.
Computing $P(H_0)$ is #P-hard in size database

Are the Lifted Rules Complete?

You already know:

- Inference rules: **PTIME** data complexity
- Some queries: **#P**-hard data complexity

Dichotomy Theorem for UCQ / Mon. CNF

- If lifted rules succeed, then **PTIME** query
- If lifted rules fail, then query is **#P**-hard

Lifted rules are complete for UCQ!

Why open world?

Knowledge Base Completion

Given:

Coauthor	x	y	P
	Einstein	Straus	0.7
	Erdos	Straus	0.6
	Einstein	Pauli	0.9

Learn:

0.8::Coauthor(x,y) :- Coauthor(z,x) \wedge Coauthor(z,y).

Complete:

x	y	P
Straus	Pauli	0.504
...

Bayesian Learning Loop

Bayesian view on learning:

1. Prior belief:

$$P(\text{Coauthor}(\text{Straus}, \text{Pauli})) = 0.01$$

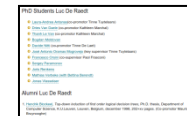
2. Observe page

$$P(\text{Coauthor}(\text{Straus}, \text{Pauli}) \mid \text{Screenshot 1}) = 0.2$$



3. Observe page

$$P(\text{Coauthor}(\text{Straus}, \text{Pauli}) \mid \text{Screenshot 2}, \text{Screenshot 1}) = 0.3$$



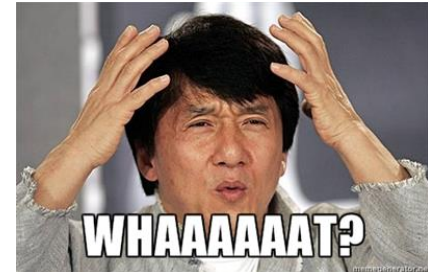
Principled and sound reasoning!

Problem: Broken Learning Loop

Bayesian view on learning:

1. Prior belief:

$$P(\text{Coauthor}(\text{Straus}, \text{Pauli})) = 0$$



2. Observe page

$$P(\text{Coauthor}(\text{Straus}, \text{Pauli} \mid \text{Screenshot of a page})) = 0.2$$

3. Observe page

$$P(\text{Coauthor}(\text{Straus}, \text{Pauli} \mid \text{Screenshot of a page}, \text{Screenshot of another page})) = 0.3$$

This is mathematical nonsense!

What we'd like to do...

$\exists x \text{ Coauthor}(\text{Einstein}, x) \wedge \text{ Coauthor}(\text{Erdos}, x)$



Ernst Straus



Kristian Kersting, ...



Justin Bieber, ...

Open World DB

- What if fact missing?
- Probability 0 for:

Coauthor

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
...

$$Q1 = \exists x \text{ Coauthor}(\text{Einstein}, x) \wedge \text{Coauthor}(\text{Erdos}, x)$$

$$Q2 = \exists x \text{ Coauthor}(\text{Bieber}, x) \wedge \text{Coauthor}(\text{Erdos}, x)$$

$$Q3 = \text{Coauthor}(\text{Einstein}, \text{Straus}) \wedge \text{Coauthor}(\text{Erdos}, \text{Straus})$$

$$Q4 = \text{Coauthor}(\text{Einstein}, \text{Bieber}) \wedge \text{Coauthor}(\text{Erdos}, \text{Bieber})$$

$$Q5 = \text{Coauthor}(\text{Einstein}, \text{Bieber}) \wedge \neg \text{Coauthor}(\text{Einstein}, \text{Bieber})$$

Intuition

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
...

$$Q1 = \exists x \text{ Coauthor}(\text{Einstein}, x) \wedge \text{Coauthor}(\text{Erdos}, x)$$

$$Q2 = \exists x \text{ Coauthor}(\text{Bieber}, x) \wedge \text{Coauthor}(\text{Erdos}, x)$$

$$Q3 = \text{Coauthor}(\text{Einstein}, \mathbf{\text{Straus}}) \wedge \text{Coauthor}(\text{Erdos}, \mathbf{\text{Straus}})$$

$$Q4 = \text{Coauthor}(\text{Einstein}, \mathbf{\text{Bieber}}) \wedge \text{Coauthor}(\text{Erdos}, \mathbf{\text{Bieber}})$$

$$Q5 = \text{Coauthor}(\text{Einstein}, \mathbf{\text{Bieber}}) \wedge \neg \text{Coauthor}(\mathbf{\text{Einstein}}, \mathbf{\text{Bieber}})$$

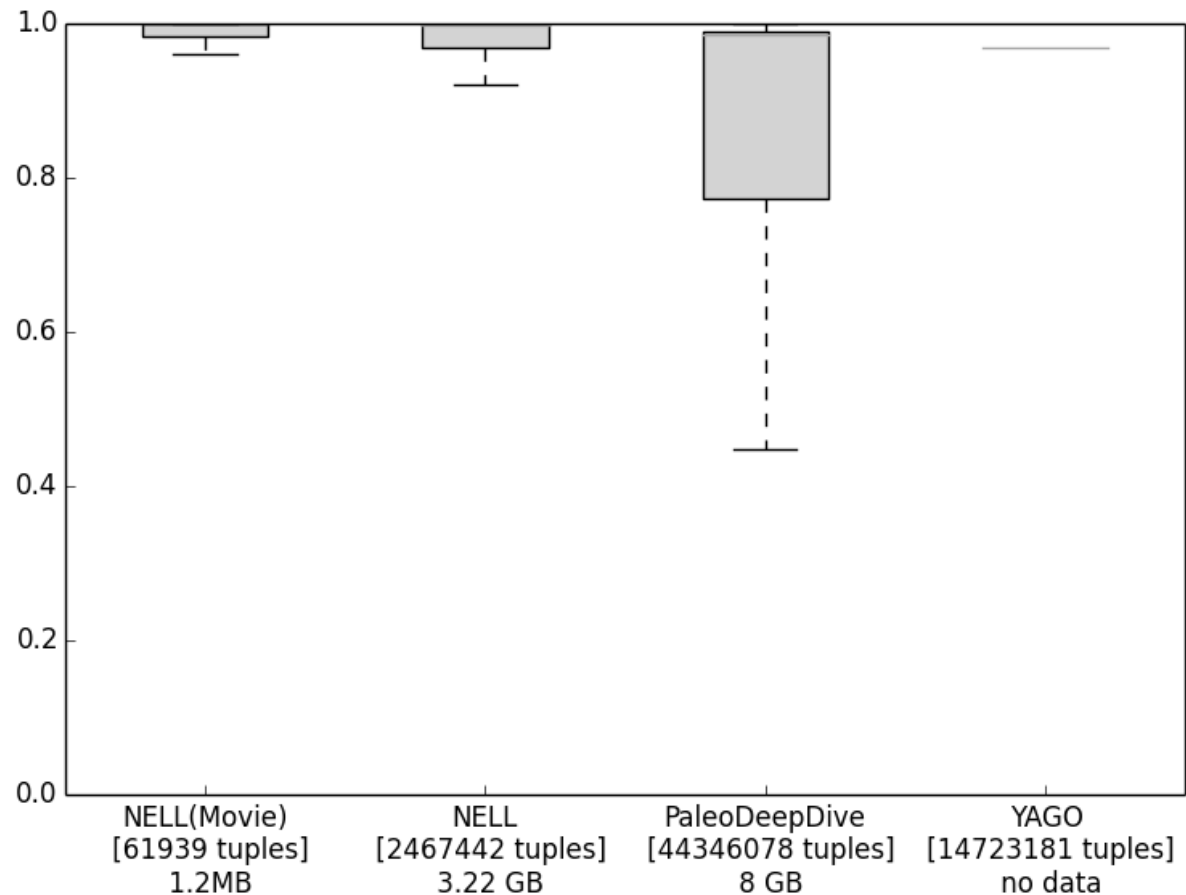
We know for sure that $P(Q1) \geq P(Q3)$, $P(Q1) \geq P(Q4)$

and $P(Q3) \geq P(Q5)$, $P(Q4) \geq P(Q5)$ because $P(Q5) = 0$.

We have strong evidence that $P(Q1) \geq P(Q2)$.

Problem: Curse of Superlinearity

- Reality is worse!
- Tuples are **intentionally** missing!



Problem: Curse of Superlinearity

Sibling

x	y	P
...

Facebook scale

⇒ 200 Exabytes of data”

All Google storage is 2 exabytes...

Randall Munroe. Google’s datacenters on punch cards, 2015.



Problem: Model Evaluation

Given:

Coauthor	x	y	P
	Einstein	Straus	0.7
	ErDOS	Straus	0.6
	Einstein	Pauli	0.9

Learn:

0.8::Coauthor(x,y) :- Coauthor(z,x) \wedge Coauthor(z,y).

OR

0.6::Coauthor(x,y) :- Affiliation(x,z) \wedge Affiliation(y,z).

What is the likelihood, precision, accuracy, ...?

Open-World Prob. Databases

Intuition: tuples can be added with $P < \lambda$

$Q2 = \text{Coauthor}(\text{Einstein}, \mathbf{\text{Straus}}) \wedge \text{Coauthor}(\text{Erdos}, \mathbf{\text{Straus}})$

$$0.7 * \lambda \geq P(Q2) \geq 0$$

Coauthor

X	Y	P
Einstein	Straus	0.7
Einstein	Pauli	0.9
Erdos	Renyi	0.7
Kersting	Natarajan	0.8
Luc	Paol	0.1
...

Coauthor

X	Y	P
Einstein	Straus	0.7
Einstein	Pauli	0.9
Erdos	Renyi	0.7
Kersting	Natarajan	0.8
Luc	Paol	0.1
...
Erdos	Straus	λ

***How open-world query
evaluation?***

UCQ / Monotone CNF

- Lower bound = closed-world probability
- Upper bound = probability after **adding all tuples** with probability λ

- Polynomial time 😊
- Quadratic blow-up 😞
- 200 exabytes ... again 😞

Closed-World Lifted Query Eval

$$Q = \exists x \exists y \text{Scientist}(x) \wedge \text{Coauthor}(x,y)$$

Decomposable \forall -Rule

$$P(Q) = 1 - \prod_{A \in \text{Domain}} (1 - P(\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y)))$$

Check independence:

$\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y)$
 $\text{Scientist}(B) \wedge \exists y \text{Coauthor}(B,y)$

$$\begin{aligned} &= 1 - (1 - P(\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y))) \\ &\quad \times (1 - P(\text{Scientist}(B) \wedge \exists y \text{Coauthor}(B,y))) \\ &\quad \times (1 - P(\text{Scientist}(C) \wedge \exists y \text{Coauthor}(C,y))) \\ &\quad \times (1 - P(\text{Scientist}(D) \wedge \exists y \text{Coauthor}(D,y))) \\ &\quad \times (1 - P(\text{Scientist}(E) \wedge \exists y \text{Coauthor}(E,y))) \\ &\quad \times (1 - P(\text{Scientist}(F) \wedge \exists y \text{Coauthor}(F,y))) \end{aligned}$$

...

Complexity PTIME

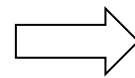
Closed-World Lifted Query Eval

$$Q = \exists x \exists y \text{Scientist}(x) \wedge \text{Coauthor}(x,y)$$

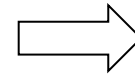
$$P(Q) = 1 - \prod_{A \in \text{Domain}} (1 - P(\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y)))$$

$$\begin{aligned} &= 1 - (1 - P(\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y)) \\ &\quad \times (1 - P(\text{Scientist}(B) \wedge \exists y \text{Coauthor}(B,y)) \\ &\quad \times (1 - P(\text{Scientist}(C) \wedge \exists y \text{Coauthor}(C,y)) \\ &\quad \times (1 - P(\text{Scientist}(D) \wedge \exists y \text{Coauthor}(D,y)) \\ &\quad \times (1 - P(\text{Scientist}(E) \wedge \exists y \text{Coauthor}(E,y)) \\ &\quad \times (1 - P(\text{Scientist}(F) \wedge \exists y \text{Coauthor}(F,y)) \end{aligned}$$

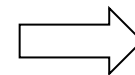
...



No supporting facts
in database!



Probability 0 in closed world



Ignore these queries!

Complexity linear time!

Open-World Lifted Query Eval

$$Q = \exists x \exists y \text{Scientist}(x) \wedge \text{Coauthor}(x,y)$$

$$P(Q) = 1 - \prod_{A \in \text{Domain}} (1 - P(\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y)))$$

$$= 1 - (1 - P(\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y)))$$

$$\times (1 - P(\text{Scientist}(B) \wedge \exists y \text{Coauthor}(B,y)))$$

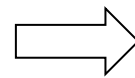
$$\times (1 - P(\text{Scientist}(C) \wedge \exists y \text{Coauthor}(C,y)))$$

$$\times (1 - P(\text{Scientist}(D) \wedge \exists y \text{Coauthor}(D,y)))$$

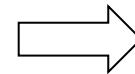
$$\times (1 - P(\text{Scientist}(E) \wedge \exists y \text{Coauthor}(E,y)))$$

$$\times (1 - P(\text{Scientist}(F) \wedge \exists y \text{Coauthor}(F,y)))$$

...



No supporting facts
in database!



Probability p in closed world

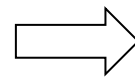
Complexity PTIME!

Open-World Lifted Query Eval

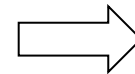
$$Q = \exists x \exists y \text{Scientist}(x) \wedge \text{Coauthor}(x,y)$$

$$P(Q) = 1 - \prod_{A \in \text{Domain}} (1 - P(\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y)))$$

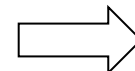
$$\begin{aligned} &= 1 - (1 - P(\text{Scientist}(A) \wedge \exists y \text{Coauthor}(A,y)) \\ &\quad \times (1 - P(\text{Scientist}(B) \wedge \exists y \text{Coauthor}(B,y)) \\ &\quad \times (1 - P(\text{Scientist}(C) \wedge \exists y \text{Coauthor}(C,y)) \\ &\quad \times (1 - P(\text{Scientist}(D) \wedge \exists y \text{Coauthor}(D,y)) \\ &\quad \times (1 - P(\text{Scientist}(E) \wedge \exists y \text{Coauthor}(E,y)) \\ &\quad \times (1 - P(\text{Scientist}(F) \wedge \exists y \text{Coauthor}(F,y)) \\ &\quad \dots \end{aligned}$$



No supporting facts
in database!



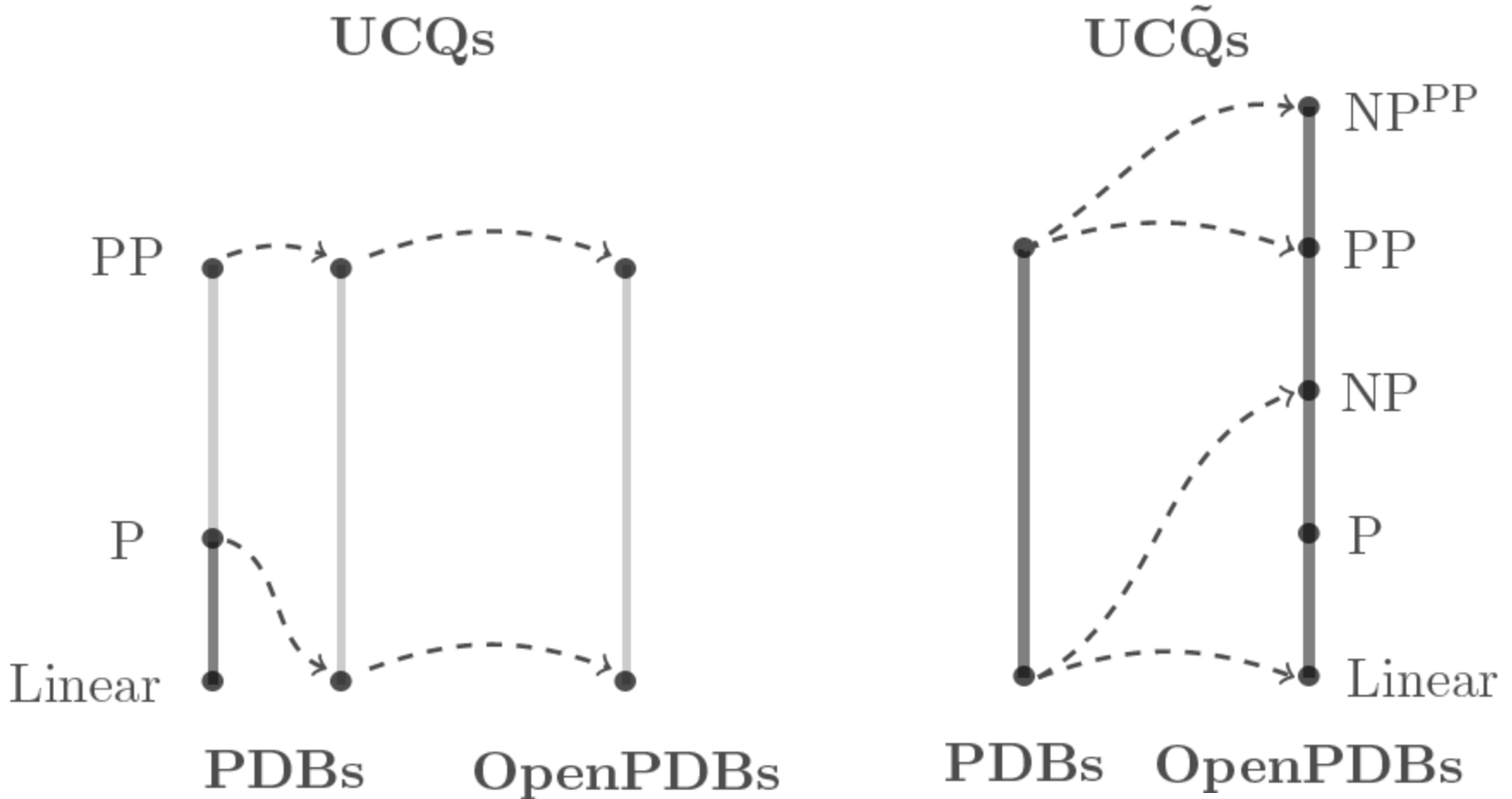
Probability p in closed world



All together, probability $(1-p)^k$
Do symmetric lifted inference

Complexity linear time!

Complexity Results



$Linear \subseteq P \subseteq NP \subseteq PP \subseteq P^{PP} \subseteq NP^{PP} \subseteq PSpace \subseteq ExpTime$

Implement PDB Query in SQL

- Convert to nested SQL recursively
- Open-world existential quantification

$$Q = \exists x P(x) \wedge Q(x)$$

```
SELECT (1.0-(1.0-pUse)*power(1.0-0.0001,(4-ct))) AS pUse
FROM
  (SELECT ior(COALESCE(pUse,0)) AS pUse,
        count(*) AS ct
   FROM SQL(conjunction))
```

0.0001 = open-world probability; 4 = # open-world query instances
ior = Independent OR aggregate function

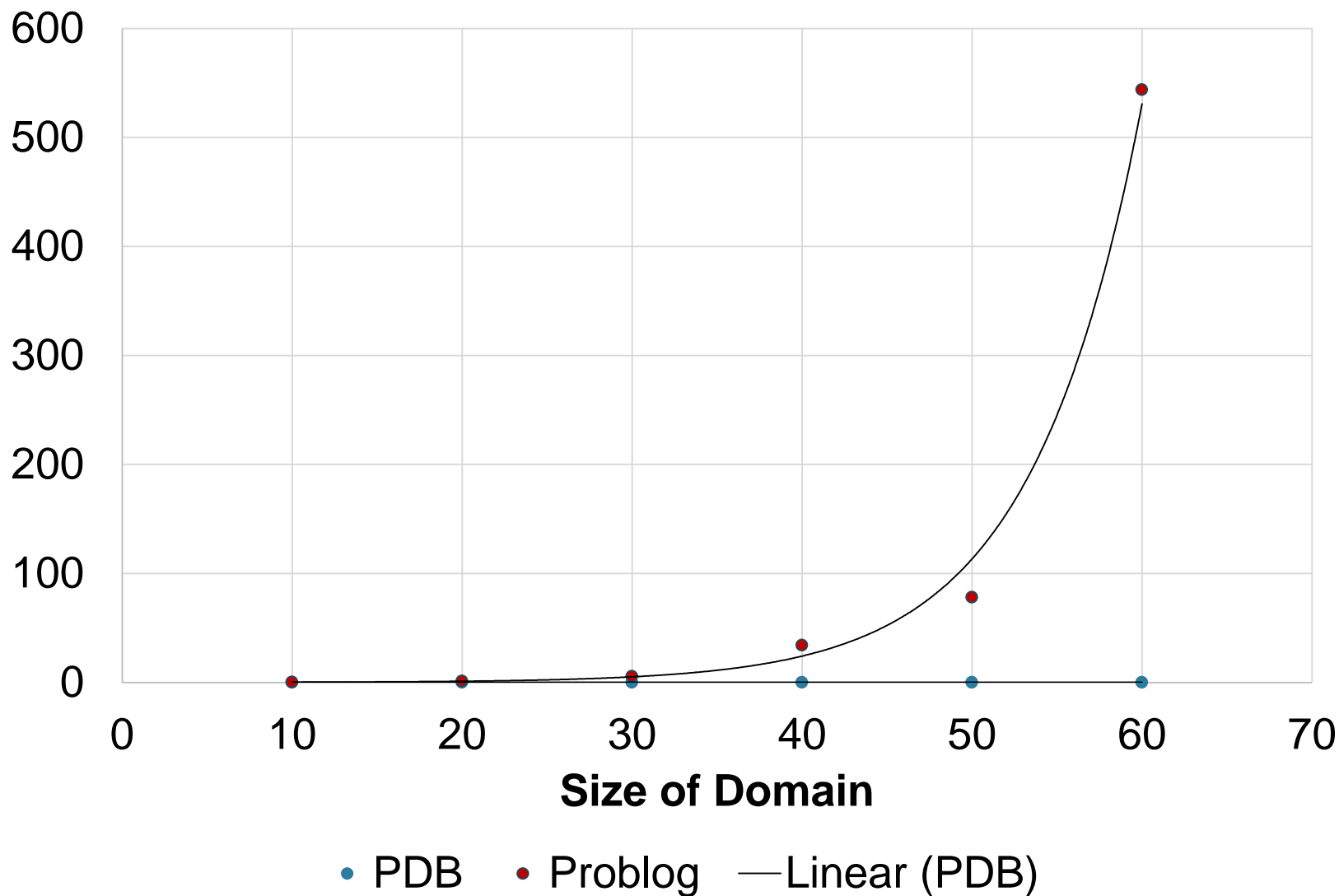
- Conjunction

```
SELECT q9.c5,
       COALESCE(q9.pUse,λ)*COALESCE(q10.pUse,λ) AS pUse
FROM
  SQL(Q(X)) OUTER JOIN SQL(P(X))
```

```
SELECT Q.v0 AS c5,
       p AS pUse
FROM Q
```

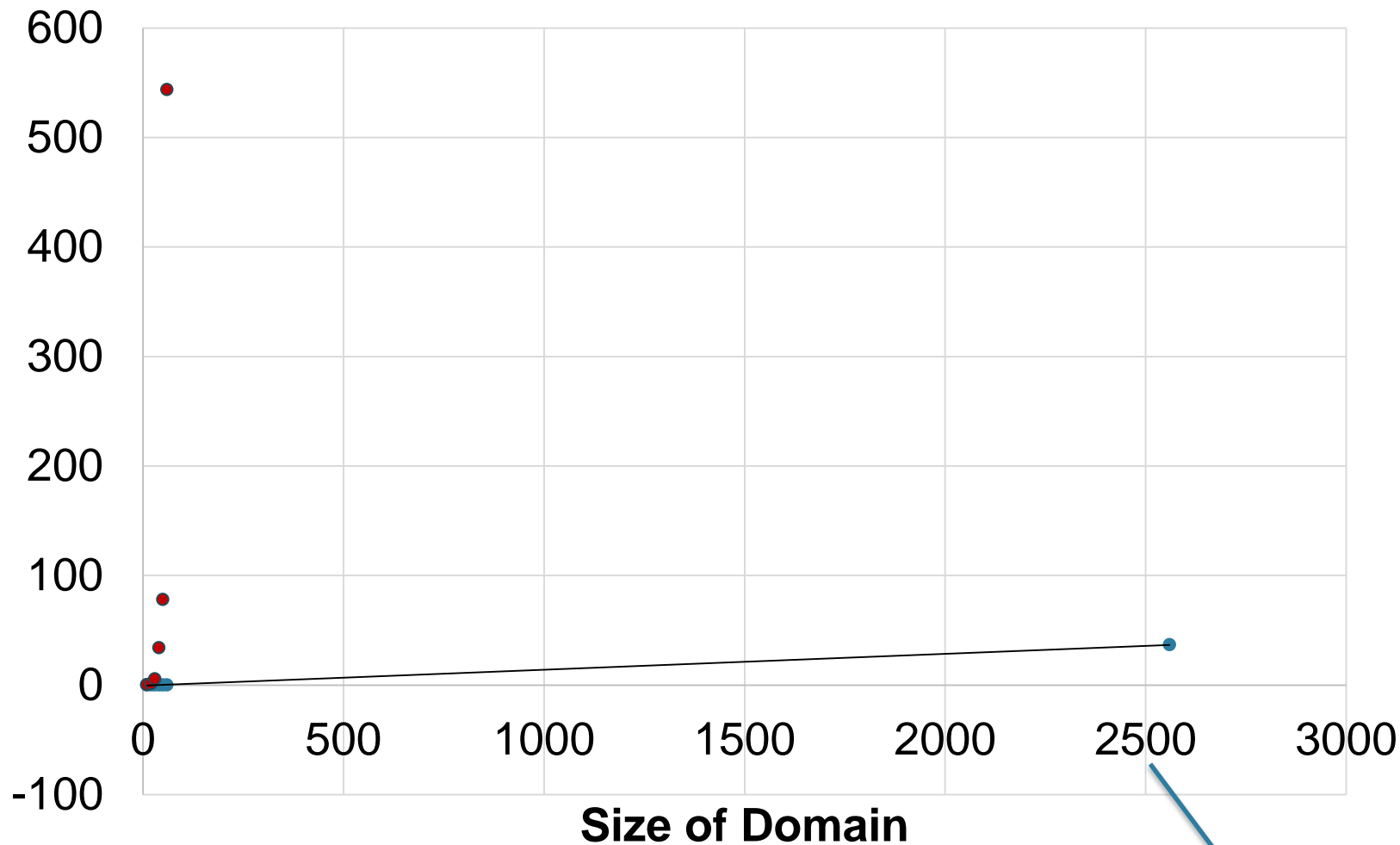
- Run as single  PostgreSQL query!

OpenPDB vs Problog Running Times (s)



Out of memory trying to run the ProbLog query with 70 constants in domain

OpenPDB vs Problog Running Times (s)



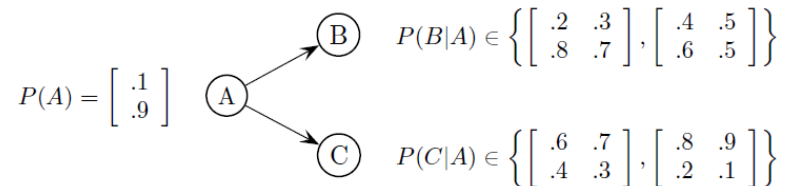
• PDB • Problog — Linear (PDB)

12.5 million
random variables!

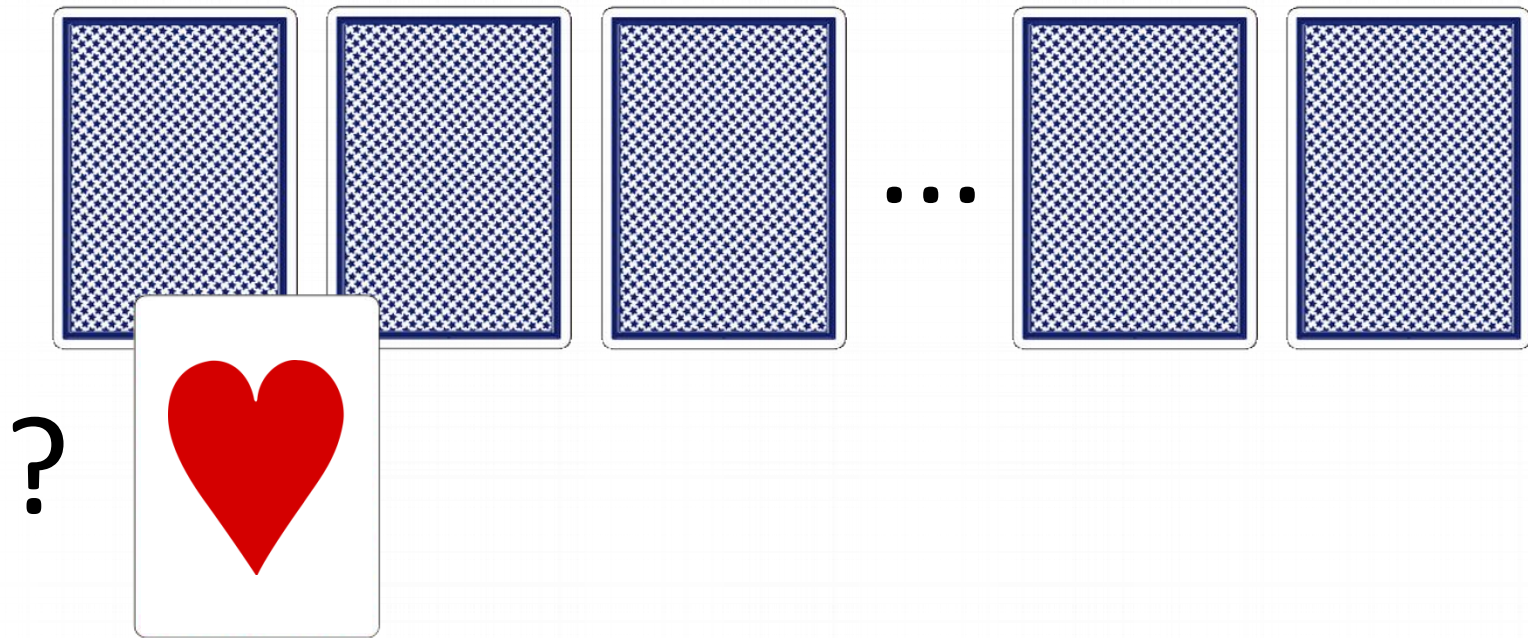
What is the broader picture?

The Broader Picture

- Statistical relational learning (e.g., Markov logic)
 - Open-domain models (BLOG)
- Probabilistic description logics
- Certain query answers in databases
- Open information extraction
- Learning from positive-only examples
- Imprecise probabilities
 - Credal sets, interval probability, qualitative uncertainty
- Credal Bayesian networks



Related Work: Lifted Probabilistic Inference



Probability that Card1 is Hearts?

$1/4$

Open-World Lifted Query Eval

$$Q = \exists x \exists y \text{ Smoker}(x) \wedge \text{Friend}(x,y)$$

$$P(Q) = 1 - \prod_{A \in \text{Domain}} (1 - P(\text{Scientist}(A) \wedge \exists y \text{ Coauthor}(A,y)))$$

$$= 1 - (1 - P(\text{Scientist}(A) \wedge \exists y \text{ Coauthor}(A,y)))$$

$$\times (1 - P(\text{Scientist}(B) \wedge \exists y \text{ Coauthor}(B,y)))$$

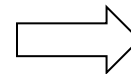
$$\times (1 - P(\text{Scientist}(C) \wedge \exists y \text{ Coauthor}(C,y)))$$

$$\times (1 - P(\text{Scientist}(D) \wedge \exists y \text{ Coauthor}(D,y)))$$

$$\times (1 - P(\text{Scientist}(E) \wedge \exists y \text{ Coauthor}(E,y)))$$

$$\times (1 - P(\text{Scientist}(F) \wedge \exists y \text{ Coauthor}(F,y)))$$

...



All together, probability $(1-p)^k$



Open-world query evaluation on empty db
= Lifted inference in AI

Conclusions

- Relational probabilistic reasoning is **frontier** and **integration** of AI, KR, ML, DB, TH, etc.
- We need
 - relational models and logic
 - probabilistic models and statistical learning
 - algorithms that scale
- Open-world data model
 - semantics makes sense
 - FREE for UCQs
 - expensive otherwise

References

- Ceylan, Ismail Ilkan, Adnan Darwiche, and Guy Van den Broeck. "Open-world probabilistic databases." Proceedings of KR (2016).
- Suciu, Dan, Dan Olteanu, Christopher Ré, and Christoph Koch. "Probabilistic databases." Synthesis Lectures on Data Management 3, no. 2 (2011): 1-180.
- Dong, Xin, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. "Knowledge vault: A web-scale approach to probabilistic knowledge fusion." In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 601-610. ACM, 2014.
- Carlson, Andrew, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr, and Tom M. Mitchell. "Toward an Architecture for Never-Ending Language Learning." In AAIL, vol. 5, p. 3. 2010.
- Niu, Feng, Ce Zhang, Christopher Ré, and Jude W. Shavlik. "DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference." VLDS 12 (2012): 25-28.

References

- Chen, Brian X. "Siri, Alexa and Other Virtual Assistants Put to the Test" The New York Times (2016).
- Dalvi, Nilesh, and Dan Suciu. "The dichotomy of probabilistic inference for unions of conjunctive queries." Journal of the ACM (JACM) 59, no. 6 (2012): 30.
- De Raedt, Luc, Anton Dries, Ingo Thon, Guy Van den Broeck, and Mathias Verbeke. "Inducing probabilistic relational rules from probabilistic examples." In Proceedings of the 24th International Conference on Artificial Intelligence, pp. 1835-1843. AAAI Press, 2015.
- Van den Broeck, Guy. "Towards high-level probabilistic reasoning with lifted inference." AAAI Spring Symposium on KRR (2015).
- Niepert, Mathias, and Guy Van den Broeck. "Tractability through exchangeability: A new perspective on efficient probabilistic inference." AAAI (2014).
- Van den Broeck, Guy. "On the completeness of first-order knowledge compilation for lifted probabilistic inference." In Advances in Neural Information Processing Systems, pp. 1386-1394. 2011.

References

- Van den Broeck, Guy, Wannes Meert, and Adnan Darwiche. "Skolemization for weighted first-order model counting." In Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR). 2014.
- Gribkoff, Eric, Guy Van den Broeck, and Dan Suci. "Understanding the complexity of lifted inference and asymmetric weighted model counting." UAI, 2014.
- Beame, Paul, Guy Van den Broeck, Eric Gribkoff, and Dan Suci. "Symmetric weighted first-order model counting." In Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, pp. 313-328. ACM, 2015.
- Chavira, Mark, and Adnan Darwiche. "On probabilistic inference by weighted model counting." Artificial Intelligence 172.6 (2008): 772-799.
- Sang, Tian, Paul Beame, and Henry A. Kautz. "Performing Bayesian inference by weighted model counting." AAI. Vol. 5. 2005.

References

- Van den Broeck, Guy, Nima Taghipour, Wannes Meert, Jesse Davis, and Luc De Raedt. "Lifted probabilistic inference by first-order knowledge compilation." In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, pp. 2178-2185. AAAI Press/International Joint Conferences on Artificial Intelligence, 2011.
- Van den Broeck, Guy. Lifted inference and learning in statistical relational models. Diss. Ph. D. Dissertation, KU Leuven, 2013.
- Gogate, Vibhav, and Pedro Domingos. "Probabilistic theorem proving." UAI (2011).

References

- Belle, Vaishak, Andrea Passerini, and Guy Van den Broeck. "Probabilistic inference in hybrid domains by weighted model integration." Proceedings of 24th International Joint Conference on Artificial Intelligence (IJCAI). 2015.
- Belle, Vaishak, Guy Van den Broeck, and Andrea Passerini. "Hashing-based approximate probabilistic inference in hybrid domains." In Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI). 2015.
- Fierens, Daan, Guy Van den Broeck, Joris Renkens, Dimitar Shterionov, Bernd Gutmann, Ingo Thon, Gerda Janssens, and Luc De Raedt. "Inference and learning in probabilistic logic programs using weighted boolean formulas." Theory and Practice of Logic Programming 15, no. 03 (2015): 358-401.