SPROUT²: A Squared Query Engine for Uncertain Web Data

SPROUT² — Structured Queries over Unstructured Date

Challenges

- Inherently uncertain Web data: NELL, Google Squared
- Scalable evaluation of any relational algebra query on
- expressive uncertain databases of Web data

Contributions

- Database wrappers for Web data
- Probabilistic databases from Google Squared, NELL
- Deterministic databases from Google Fusion
- Scalable query processing for any relational algebra quer
- Exact/Approximate evaluation for tractable/hard queries
- Done by the SPROUT query engine (part of MayBMS)
- Graphical user interface in SPROUT²'s for selecting data sources, composing queries, and browsing results

Uncertain Data on the Web

- Traditional databases: Data must adhere to rigid structure and must be complete and trustworthy.
- Internet applications witness an unprecedented shift towards unstructured and user-generated content.
- Example scenario: Business Intelligence
 - Extract valuable business information from sources suc as twitter feeds, blogs, or email messages.
 - Join those data with offline databases, e.g. on products
- ... to obtain early feedback about a product's quality.

Selected Publications on SPROUT and SPROUT²

- ► SPROUT²: A Squared Query Engine for Uncertain Web Data. R. Fink, D. Olteanu, S. Rath. In SIGMOD, 2011.
- On the Optimal Approximation of Queries Using Tractable Propositional Languages. R. Fink, D. Olteanu. In ICDT, 2011.
- Providing Support for Full Relational Algebra in Probabilistic Databases. R. Fink, D. Olteanu, S. Rath. In ICDE, 2011.
- Approximate Confidence Computation for Probabilistic Databases. D. Olteanu, J. Huang, C. Koch. In *ICDE*, 2010.
- Secondary-Storage Confidence Computation for Conjunctive Queries with Inequalities. D. Olteanu, J. Huang. In SIGMOD, 2009
- SPROUT: Lazy vs. Eager Query Plans for Tuple-independent Probabilistic Databases. Olteanu, Huang, Koch. In ICDE, 2009.
- Using OBBDs for Efficient Query Evaluation on Probabilistic Databases. D. Olteanu, J. Huang. In SUM, 2008.

New Book on Probabilistic Databases

Probabilistic Databases. D. Suciu, D. Olteanu, C. Ré, C. Koch. May 2011. Morgan & Claypool Publishers.

Robert Fink¹, Andrew Hogue², Dan Olteanu¹, Swaroop Rath¹

²Google Inc., New York, NY, USA ¹Department of Computer Science, University of Oxford, Oxford, OX1 3QD, UK

ta	Query Answering in SPROUT
	Query evaluation in two steps (logically distinct but pose 1. Compute the tuples in the query result together with The lineage of a result tuple <i>t</i> is a propositional formula over the which input tuples must be present in order for the query to return 2. Compute the probabilities of result tuples by increment Independent or (ϕ and ψ are syntactically independent) P($\phi \lor \psi$ Shannon expansion (<i>x</i> is a variable in ϕ) P($\phi \models \psi$
ery es) a	Example: Relational division query "Which suppliesSupplier SProduct PSupplier of all products: $S \div P$ $\overline{Supplier Product E}$ $\overline{Product E}$ $\overline{Product E}$ $\overline{Supplier E}$ 1 2 x_2 2 y_2 2 1 7 x_3 2 1 x_4 2 5 x_5 $P((x_4 \lor x_5)(y_1 \to x_4)(\neg y_2)) = P(\neg y_2)[P(y_1)P(x_4) + P(\neg y_1)[1])$
	Novel Techniques for Exact and Approximate Prob
re	 Incremental decomposition of lineage into d-trees us After each decomposition step, compute rough lower probabilities of the residual formulas at the leafs of the Approach 1: Lower bound is the largest probability of a claus probabilities of all clauses in Φ.
ich ts	 Approach 2: Compute read-once formulas, whose probability Using the bounds at the leaves, compute lower and it Stop when the desired precision is reached or the tir Underlying idea: Leaves deeper in the d-tree contribution mass, hence a good approximation can be found quite
	 Complete decomposition Corresponds to exact probability computation Can be done in polynomial time for tractable query 8 relational algebra queries without repeating symbol a class of conjunctive queries with inequality (<, ≠) Partial decomposition Corresponds to approximate probability computation Applicable for hard query and data instances
9.	 Example: Efficient computation of bounds that are Left: original formula; middle: lower bound; right: up Lower/Upper bounds obtained by setting the marked
	$\begin{array}{c} & & & & & \\ & & & &$
	$(x_1y_1 \lor \neg x_2y_2)[x_1(y_1 \lor y_3) \lor x_3(y_4 \lor y_5)]$ Lower bound $(x_1y_1 \lor \neg x_2y_2)x_3(y_4 \lor y_5)$



ssibly intertwined):

their lineage

he tuples in the input database and says eturn *t*.

ental lineage compilation:

 $\Psi(\Psi) = 1 - (1 - P(\Phi))(1 - P(\Psi))$ $(\Psi) = P(\Phi) \cdot P(\Psi)$ $= P(x)P(\Phi|_{x}) + P(\neg x)P(\Phi|_{\neg x})$

r stocks all products?"

 $(X_4 \lor X_5)(Y_1 \rightarrow X_4)(\neg Y_2)$ $(X_4 \lor X_5)(Y_1 \to X_4)$ $\neg y_2$

 $\neg y_1(x_4 \lor x_5)$

bability Computation

 $-(1 - P(x_4))(1 - P(x_5))$

ising the above three rules er and upper bounds on the the decomposition tree use in Φ ; Upper bound is the sum of

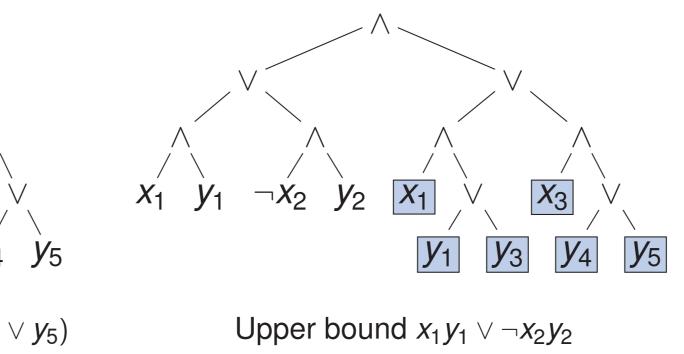
ities represent lower and upper bounds. upper bounds for the whole lineage me budget is exhausted bute little to the overall probability lickly

& data instances ools and with read-once lineage) joins

n with error guarantees

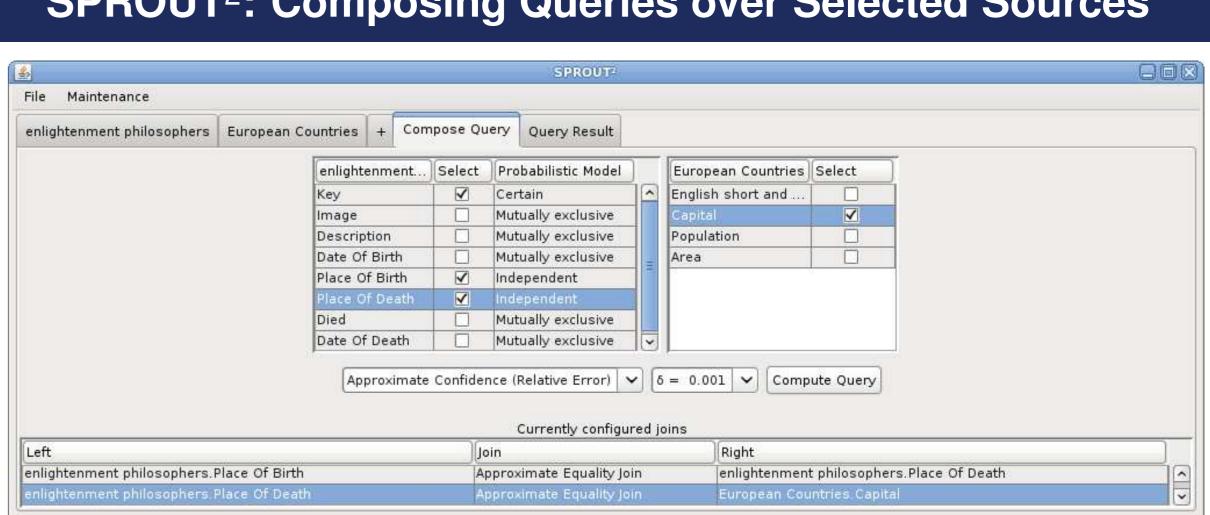
read-once formulas

per bound d literals to false/true



SPROUT²: Selecting Web or Offline Data Sources





SPROUT²: Browsing Ranked Query Results

SPROUT2							
File Ma	intena	nce					
	enlightenment philosophers						
+		Compose					
enlighten	me)	enlightenme	enlightenme	967574	4. c		
Thomas		London	London	London			
Voltaine		Paris, France	Rans: France	Paris			
Denis Did	erot	Langres, Fra	Paris, France	Paris			
Charles d	le S	Bordeaux, Fr	Paris, France	Paris			
Wolfgang	Am	Salzburg, Au	Vienna, Austria	Vienna			
Voltaire		Paris, France	Paris	Paris			
Baruch S	pinoza	Amsterdam,	Hague, Nethe	Prague	22		
Voltaire		Paris	Paris, France	Paris			
Wolfgang Am		Vienna, Austria	Vienna	Vienna			
Voltaire		Paris, France	Paris, France	Amsterda		Amsterd	
Baruch S	pinoza	Amsterdam,	Hague, Nethe	Amster	da		
Wolfgang	Am	Vienna, Austria	Vienna, Austria	Vienna			

Publicly-available Prototype

- 8.4 backend
- maybms.sourceforge.net

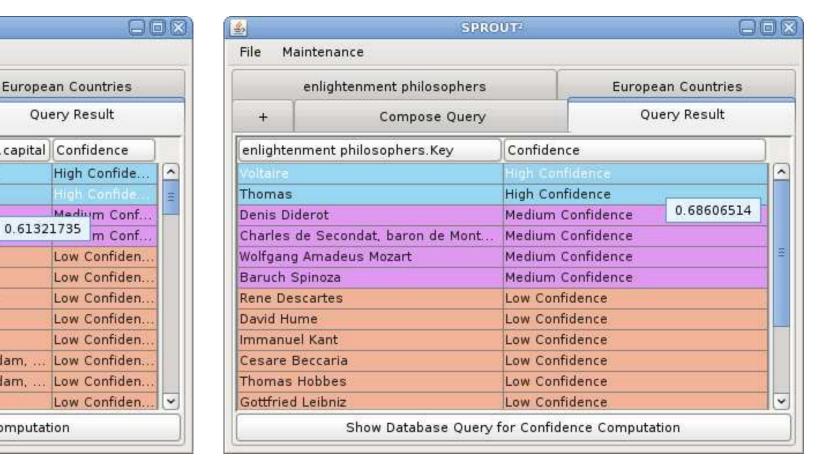
DEPARTMENT OF COMPUTER SCIENCE



	1	~]	Retri	eve Google Square		
Date Of Birth Place		Place Of Birth	ce Of Birth Place Of Death Died		Date Of Death	
	21 November 1694	Paris, France	Paris, France	May 30, 1778	1778-05-30	
influe	29 August 1632	Wrington, Somerse	Essex, England	1704-10-28	1704-10-28	
(7 Ma	1711-04-26	Edinburgh, Scotland	Edinburgh, Scot	Paris, France high conf	idence)8-25	
nt wa	April 22, 1724	Kaliningrad, Russia	Kaliningrad, Rus	France low confi	dence)2-12	
s de,	January 18, 1689	Bordeaux, France	Paris, France	Paris low confi	dence ary 10, 1755	5
t (Oct	5 October 1713	Langres, France	Paris, France	1/84-07-31	1784-07-31	
Rous	June 28, 1712		Malone, NY	7/2/1778	June 28, 1712	
nt phil	April 5, 1588	Malmesbury, Wilts	Derbyshire, Engl	and 1679-12-04	December 4, 1679	9
(bapti	1991-04-29	Kirkcaldy, Scotland	Edinburgh, Scotland 1790-07-17		1790-07-17	
n" Pai	1737-01-29	Thetford, England	New York City	8-Jun-1809	June 8, 1809	
alist p	November 24, 1632	Amsterdam	Hague, Netherlar	nds 1677-02-21	1677-02-21	
tes Fr	March 31, 1596	La Haye, France	Stockholm, Swee	len 11 February 1650	February 11, 1650)
ielm L	1 Iulv 1646	Leipzig, Germany	Hanover, Germa	nv 14 November 1716	14 November 171	6
S	how Database Queri	es for Caching of this	Table			
	18					
	(closed	11.000			1	100
	SPRO	1011-			2.lg	

	✓ Retrieve Google Fusion Table		
	Population	Area	
	2986952	28,748 km2(11,100 sq mi)	
a	84525	468 km2(181 sq mi)	
	2966802	29,743 km2(11,484 sq mi)[n 1]	
	8214160	83,871 km2(32,383 sq mi)	
	8303512	86,600 km2(33,436 sq mi)[n 1]	
	9612632	207,600 km2(80,155 sq mi)	
	10423493	30,528 km2(11,787 sq mi)	1
	3843126	51,129 km2(19,741 sq mi)	
	7148785	110,879 km2(42,811 sq mi)	
	4486881	56,594 km2(21,851 sq mi)	1
	1102677	9,251 km2(3,572 sq mi)[n 1]	
	10201777	78,867 km2(30,451 sq mi)	
	5515575	43.094 km2(16.639 sa mi)[n 1]	

SPROUT²: Composing Queries over Selected Sources



SPROUT is implemented as an extension of PostgreSQL

Public version downloadable with MayBMS from CVS at