Aggregation in Probabilistic Databases

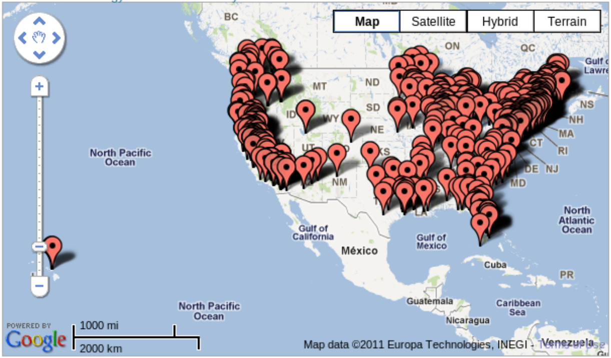
Robert Fink, University of Oxford
rf@robertfink.de

Télécom ParisTech
DBWeb Seminar
January 26th 2012

Outline

Biomass Plants in the US

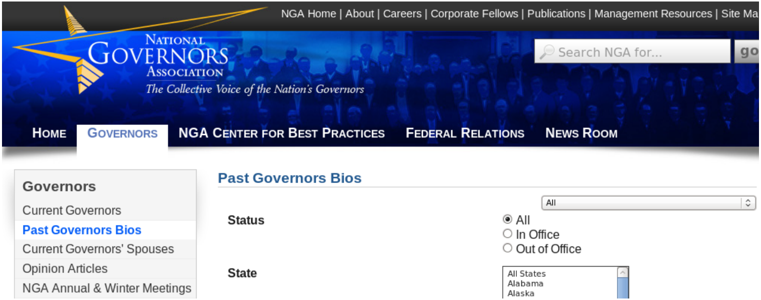
✚ Add a new Energy Generation Facility



AES Mendota Biomass Facility
APS Biomass I Biomass Facility
Aberdeen Biomass Facility
Acme Landfill Biomass Facility
Adrian Energy Associates LLC Biomass Facility
Agrilectric Power Partners Ltd Biomass Facility
Al Turi Biomass Facility
Alabama Pine Pulp Biomass Facility
Albany Landfill Gas Utilization Project Biomass Facility
Alexandria Biomass Facility
Altamont Gas Recovery Biomass Facility
American Canyon Power Plant Biomass Facility
American Ref-Fuel of Delaware Valley

Name	Alabama Pine Pulp Biomass Facility
Facility	Alabama Pine Pulp
Sector	Biomass
Location	Monroe County, Alabama
Coordinates	31.5119068°, -87.460397° Display map
Generating Capacity (MW)	32.085
Commercial Online Date	1991

Governors in US States



NGA Home | About | Careers | Corporate Fellows | Publications | Management Resources | Site Ma

National Governors Association

The Collective Voice of the Nation's Governors

Governors

- Current Governors
- Past Governors Bios
- Current Governors' Spouses
- Opinion Articles
- NGA Annual & Winter Meetings

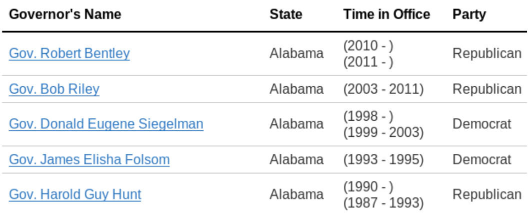
Past Governors Bios

All

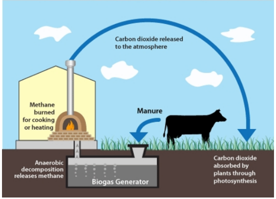
Status ● All ○ In Office ○ Out of Office

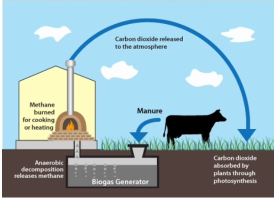
State

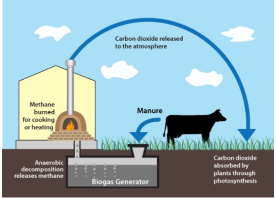
All States
Alabama
Alaska



Governor's Name	State	Time in Office	Party
Gov. Robert Bentley	Alabama	(2010 -) (2011 -)	Republican
Gov. Bob Riley	Alabama	(2003 - 2011)	Republican
Gov. Donald Eugene Siegelman	Alabama	(1998 -) (1999 - 2003)	Democrat
Gov. James Elisha Folsom	Alabama	(1993 - 1995)	Democrat
Gov. Harold Guy Hunt	Alabama	(1990 -) (1987 - 1993)	Republican


Carbon dioxide released to the atmosphere
Methane burned for cooking or heating
Manure
Anaerobic decomposition releases methane
Biogas Generator
Carbon dioxide absorbed by plants through photosynthesis

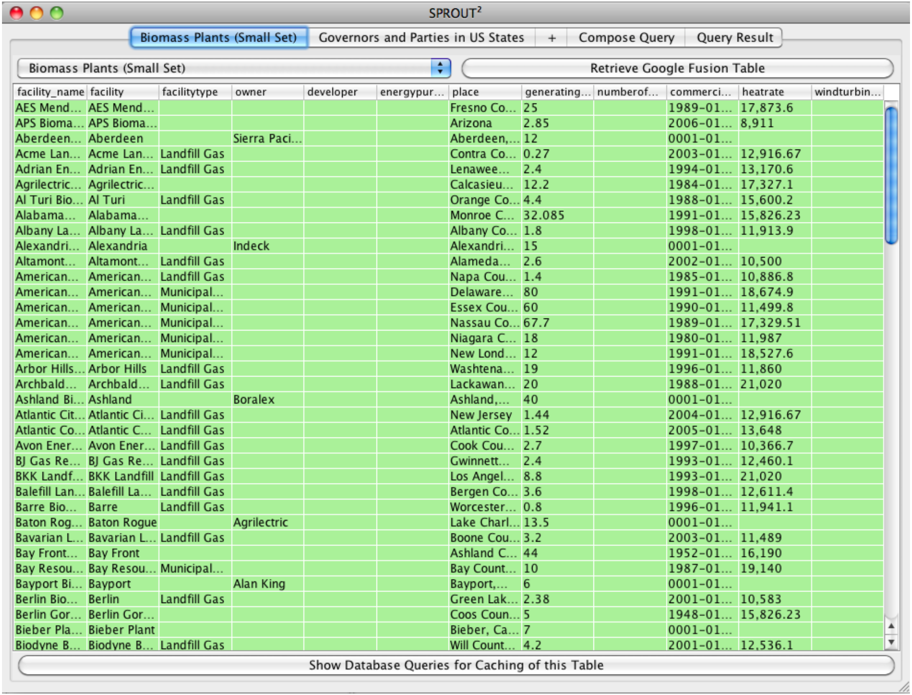

Carbon dioxide released
to the atmosphere
Methane
burned
for cooking
or heating
Manure
Anaerobic
decomposition
releases methane
Biogas Generator
Carbon dioxide
absorbed by
plants through
photosynthesis


Carbon dioxide released to the atmosphere
Methane burned for cooking or heating
Manure
Anaerobic decomposition releases methane
Biogas Generator
Carbon dioxide absorbed by plants through photosynthesis

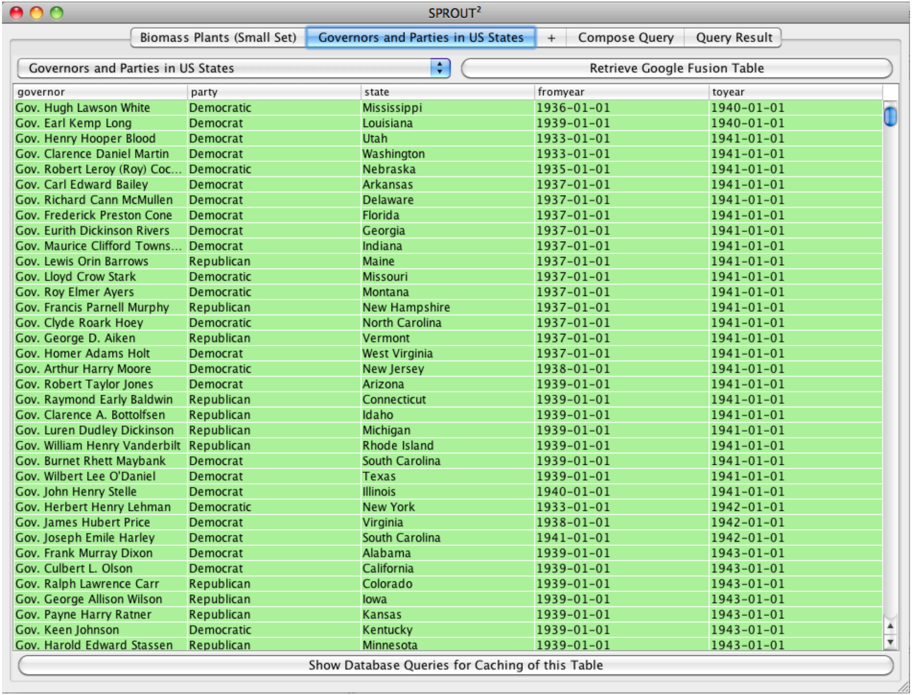
?

Who is responsible for a larger capacity of biogas plants, Democrats or Republicans?

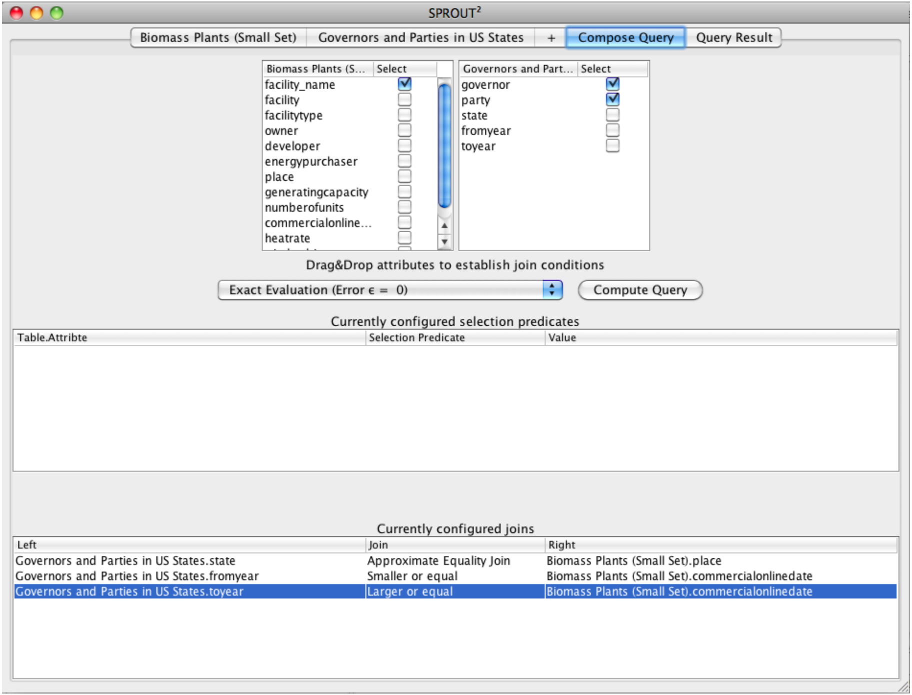
?


SPROUT²
Biomass Plants (Small Set)
Governors and Parties in US States
+
Compose Query
Query Result
Biomass Plants (Small Set)
Retrieve Google Fusion Table
Show Database Queries for Caching of this Table

facility_name	facility	facilitytype	owner	developer	energypur...	place	generating...	numberof...	commerci...	heatrate	windturbin...
AES Mend...	AES Mend...					Fresno Co...	25		1989-01...	17,873.6	
APS Bioma...	APS Bioma...					Arizona	2.85		2006-01...	8,911	
Aberdeen...	Aberdeen		Sierra Paci...			Aberdeen,...	12		0001-01...		
Acme Lan...	Acme Lan...	Landfill Gas				Contra Co...	0.27		2003-01...	12,916.67	
Adrian En...	Adrian En...	Landfill Gas				Lenawee...	2.4		1994-01...	13,170.6	
Agrilectric...	Agrilectric...					Calcasieu...	12.2		1984-01...	17,327.1	
Al Turi Bio...	Al Turi	Landfill Gas				Orange Co...	4.4		1988-01...	15,600.2	
Alabama...	Alabama...					Monroe C...	32.085		1991-01...	15,826.23	
Albany La...	Albany La...	Landfill Gas				Albany Co...	1.8		1998-01...	11,913.9	
Alexandri...	Alexandria		Indeck			Alexandri...	15		0001-01...		
Altamont...	Altamont...	Landfill Gas				Alameda...	2.6		2002-01...	10,500	
American...	American...	Landfill Gas				Napa Cou...	1.4		1985-01...	10,886.8	
American...	American...	Municipal...				Delaware...	80		1991-01...	18,674.9	
American...	American...	Municipal...				Essex Cou...	60		1990-01...	11,499.8	
American...	American...	Municipal...				Nassau Co...	67.7		1989-01...	17,329.51	
American...	American...	Municipal...				Niagara C...	18		1980-01...	11,987	
American...	American...	Municipal...				New Lond...	12		1991-01...	18,527.6	
Arbor Hills...	Arbor Hills	Landfill Gas				Washtena...	19		1996-01...	11,860	
Archbald...	Archbald...	Landfill Gas				Lackawan...	20		1988-01...	21,020	
Ashland Bi...	Ashland		Boralex			Ashland,...	40		0001-01...		
Atlantic Cit...	Atlantic Ci...	Landfill Gas				New Jersey	1.44		2004-01...	12,916.67	
Atlantic Co...	Atlantic C...	Landfill Gas				Atlantic Co...	1.52		2005-01...	13,648	
Avon Ener...	Avon Ener...	Landfill Gas				Cook Cou...	2.7		1997-01...	10,366.7	
BJ Gas Re...	BJ Gas Re...	Landfill Gas				Gwinnett...	2.4		1993-01...	12,460.1	
BKK Landf...	BKK Landfill	Landfill Gas				Los Angel...	8.8		1993-01...	21,020	
Balefill Lan...	Balefill La...	Landfill Gas				Bergen Co...	3.6		1998-01...	12,611.4	
Barre Bio...	Barre	Landfill Gas				Worcester...	0.8		1996-01...	11,941.1	
Baton Rog...	Baton Rogue		Agrilectric			Lake Charl...	13.5		0001-01...		
Bavarian L...	Bavarian L...	Landfill Gas				Boone Cou...	3.2		2003-01...	11,489	
Bay Front...	Bay Front					Ashland C...	44		1952-01...	16,190	
Bay Resou...	Bay Resou...	Municipal...				Bay Count...	10		1987-01...	19,140	
Bayport Bi...	Bayport		Alan King			Bayport,...	6		0001-01...		
Berlin Bio...	Berlin	Landfill Gas				Green Lak...	2.38		2001-01...	10,583	
Berlin Gor...	Berlin Gor...					Coos Coun...	5		1948-01...	15,826.23	
Bieber Pla...	Bieber Plant					Bieber, Ca...	7		0001-01...		
Biodyne B...	Biodyne B...	Landfill Gas				Will Count...	4.2		2001-01...	12,536.1	


SPROUT²
Biomass Plants (Small Set)
Governors and Parties in US States
+
Compose Query
Query Result
Governors and Parties in US States
Retrieve Google Fusion Table
Show Database Queries for Caching of this Table

governor	party	state	fromyear	toyear
Gov. Hugh Lawson White	Democratic	Mississippi	1936-01-01	1940-01-01
Gov. Earl Kemp Long	Democrat	Louisiana	1939-01-01	1940-01-01
Gov. Henry Hooper Blood	Democrat	Utah	1933-01-01	1941-01-01
Gov. Clarence Daniel Martin	Democrat	Washington	1933-01-01	1941-01-01
Gov. Robert Leroy (Roy) Coc...	Democratic	Nebraska	1935-01-01	1941-01-01
Gov. Carl Edward Bailey	Democrat	Arkansas	1937-01-01	1941-01-01
Gov. Richard Cann McMullen	Democrat	Delaware	1937-01-01	1941-01-01
Gov. Frederick Preston Cone	Democrat	Florida	1937-01-01	1941-01-01
Gov. Eurith Dickinson Rivers	Democrat	Georgia	1937-01-01	1941-01-01
Gov. Maurice Clifford Towns...	Democrat	Indiana	1937-01-01	1941-01-01
Gov. Lewis Orin Barrows	Republican	Maine	1937-01-01	1941-01-01
Gov. Lloyd Crow Stark	Democratic	Missouri	1937-01-01	1941-01-01
Gov. Roy Elmer Ayers	Democratic	Montana	1937-01-01	1941-01-01
Gov. Francis Parnell Murphy	Republican	New Hampshire	1937-01-01	1941-01-01
Gov. Clyde Roark Hoey	Democratic	North Carolina	1937-01-01	1941-01-01
Gov. George D. Aiken	Republican	Vermont	1937-01-01	1941-01-01
Gov. Homer Adams Holt	Democrat	West Virginia	1937-01-01	1941-01-01
Gov. Arthur Harry Moore	Democratic	New Jersey	1938-01-01	1941-01-01
Gov. Robert Taylor Jones	Democrat	Arizona	1939-01-01	1941-01-01
Gov. Raymond Early Baldwin	Republican	Connecticut	1939-01-01	1941-01-01
Gov. Clarence A. Bottolfsen	Republican	Idaho	1939-01-01	1941-01-01
Gov. Luren Dudley Dickinson	Republican	Michigan	1939-01-01	1941-01-01
Gov. William Henry Vanderbilt	Republican	Rhode Island	1939-01-01	1941-01-01
Gov. Burnet Rhett Maybank	Democrat	South Carolina	1939-01-01	1941-01-01
Gov. Wilbert Lee O'Daniel	Democrat	Texas	1939-01-01	1941-01-01
Gov. John Henry Stelle	Democrat	Illinois	1940-01-01	1941-01-01
Gov. Herbert Henry Lehman	Democratic	New York	1933-01-01	1942-01-01
Gov. James Hubert Price	Democrat	Virginia	1938-01-01	1942-01-01
Gov. Joseph Emile Harley	Democrat	South Carolina	1941-01-01	1942-01-01
Gov. Frank Murray Dixon	Democrat	Alabama	1939-01-01	1943-01-01
Gov. Culbert L. Olson	Democrat	California	1939-01-01	1943-01-01
Gov. Ralph Lawrence Carr	Republican	Colorado	1939-01-01	1943-01-01
Gov. George Allison Wilson	Republican	Iowa	1939-01-01	1943-01-01
Gov. Payne Harry Ratner	Republican	Kansas	1939-01-01	1943-01-01
Gov. Keen Johnson	Democratic	Kentucky	1939-01-01	1943-01-01
Gov. Harold Edward Stassen	Republican	Minnesota	1939-01-01	1943-01-01


SPROUT^2
Biomass Plants (Small Set)
Governors and Parties in US States
+
Compose Query
Query Result
Compute Query

Biomass Plants (S…	Select
facility_name	☑
facility	☐
facilitytype	☐
owner	☐
developer	☐
energypurchaser	☐
place	☐
generatingcapacity	☐
numberofunits	☐
commercialonline…	☐
heatrate	☐

Governors and Part…	Select
governor	☑
party	☑
state	☐
fromyear	☐
toyear	☐

Drag&Drop attributes to establish join conditions

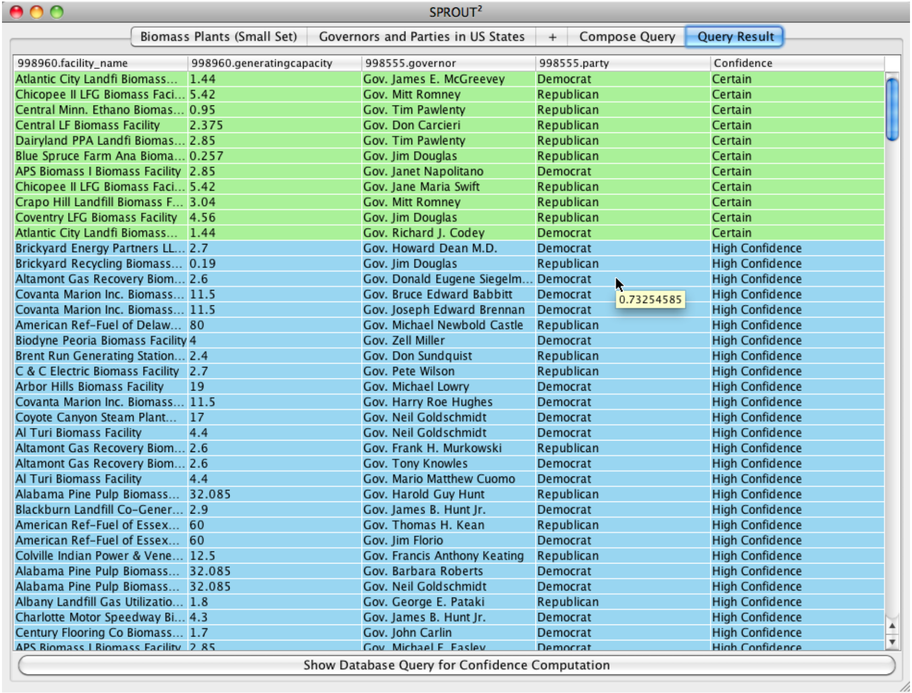
Exact Evaluation (Error ϵ = 0)

Currently configured selection predicates

Table.Attribte	Selection Predicate	Value

Currently configured joins

Left	Join	Right
Governors and Parties in US States.state	Approximate Equality Join	Biomass Plants (Small Set).place
Governors and Parties in US States.fromyear	Smaller or equal	Biomass Plants (Small Set).commercialonlinedate
Governors and Parties in US States.toyear	Larger or equal	Biomass Plants (Small Set).commercialonlinedate



SPROUT²
Biomass Plants (Small Set)
Governors and Parties in US States
+
Compose Query
Query Result
0.73254585
Show Database Query for Confidence Computation

998960.facility_name	998960.generatingcapacity	998555.governor	998555.party	Confidence
Atlantic City Landfi Biomass...	1.44	Gov. James E. McGreevey	Democrat	Certain
Chicopee II LFG Biomass Faci...	5.42	Gov. Mitt Romney	Republican	Certain
Central Minn. Ethano Biomas...	0.95	Gov. Tim Pawlenty	Republican	Certain
Central LF Biomass Facility	2.375	Gov. Don Carcieri	Republican	Certain
Dairyland PPA Landfi Biomas...	2.85	Gov. Tim Pawlenty	Republican	Certain
Blue Spruce Farm Ana Bioma...	0.257	Gov. Jim Douglas	Republican	Certain
APS Biomass I Biomass Facility	2.85	Gov. Janet Napolitano	Democrat	Certain
Chicopee II LFG Biomass Faci...	5.42	Gov. Jane Maria Swift	Republican	Certain
Crapo Hill Landfill Biomass F...	3.04	Gov. Mitt Romney	Republican	Certain
Coventry LFG Biomass Facility	4.56	Gov. Jim Douglas	Republican	Certain
Atlantic City Landfi Biomass...	1.44	Gov. Richard J. Codey	Democrat	Certain
Brickyard Energy Partners LL...	2.7	Gov. Howard Dean M.D.	Democrat	High Confidence
Brickyard Recycling Biomass...	0.19	Gov. Jim Douglas	Republican	High Confidence
Altamont Gas Recovery Biom...	2.6	Gov. Donald Eugene Siegelm...	Democrat	High Confidence
Covanta Marion Inc. Biomass...	11.5	Gov. Bruce Edward Babbitt	Democrat	High Confidence
Covanta Marion Inc. Biomass...	11.5	Gov. Joseph Edward Brennan	Democrat	High Confidence
American Ref-Fuel of Delaw...	80	Gov. Michael Newbold Castle	Republican	High Confidence
Biodyne Peoria Biomass Facility	4	Gov. Zell Miller	Democrat	High Confidence
Brent Run Generating Station...	2.4	Gov. Don Sundquist	Republican	High Confidence
C & C Electric Biomass Facility	2.7	Gov. Pete Wilson	Republican	High Confidence
Arbor Hills Biomass Facility	19	Gov. Michael Lowry	Democrat	High Confidence
Covanta Marion Inc. Biomass...	11.5	Gov. Harry Roe Hughes	Democrat	High Confidence
Coyote Canyon Steam Plant...	17	Gov. Neil Goldschmidt	Democrat	High Confidence
Al Turi Biomass Facility	4.4	Gov. Neil Goldschmidt	Democrat	High Confidence
Altamont Gas Recovery Biom...	2.6	Gov. Frank H. Murkowski	Republican	High Confidence
Altamont Gas Recovery Biom...	2.6	Gov. Tony Knowles	Democrat	High Confidence
Al Turi Biomass Facility	4.4	Gov. Mario Matthew Cuomo	Democrat	High Confidence
Alabama Pine Pulp Biomass...	32.085	Gov. Harold Guy Hunt	Republican	High Confidence
Blackburn Landfill Co-Gener...	2.9	Gov. James B. Hunt Jr.	Democrat	High Confidence
American Ref-Fuel of Essex...	60	Gov. Thomas H. Kean	Republican	High Confidence
American Ref-Fuel of Essex...	60	Gov. Jim Florio	Democrat	High Confidence
Colville Indian Power & Vene...	12.5	Gov. Francis Anthony Keating	Republican	High Confidence
Alabama Pine Pulp Biomass...	32.085	Gov. Barbara Roberts	Democrat	High Confidence
Alabama Pine Pulp Biomass...	32.085	Gov. Neil Goldschmidt	Democrat	High Confidence
Albany Landfill Gas Utilizatio...	1.8	Gov. George E. Pataki	Republican	High Confidence
Charlotte Motor Speedway Bi...	4.3	Gov. James B. Hunt Jr.	Democrat	High Confidence
Century Flooring Co Biomass...	1.7	Gov. John Carlin	Democrat	High Confidence
APS Biomass I Biomass Facility	2.85	Gov. Michael F. Easley	Democrat	High Confidence

Algebraic Expressions give rise to Random Variables

$$\begin{array}{ccc} \text{Democratic Biomass Capacity} & \geq & \text{Republican Biomass Capacity} \\ 17 + \quad 5 + \quad 9 & \geq & 8 + \quad 14 + \quad 2 \end{array}$$

Algebraic Expressions give rise to Random Variables

$$\begin{aligned} & \text{Democratic Biomass Capacity} && \geq && \text{Republican Biomass Capacity} \\ \Phi \quad = \quad & [x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 && \geq && x_4 \otimes 8 + x_5 \otimes 14 + x_6 \otimes 2] \end{aligned}$$

Algebraic Expressions give rise to Random Variables

$$\begin{aligned} & \text{Democratic Biomass Capacity} & \geq & \quad \text{Republican Biomass Capacity} \\ \Phi = & [x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 & \geq & \quad x_4 \otimes 8 + x_5 \otimes 14 + x_6 \otimes 2] \end{aligned}$$

- Assume x_i are Boolean random variables

Algebraic Expressions give rise to Random Variables

$$\begin{aligned} & & \text{Democratic Biomass Capacity} & \geq \text{Republican Biomass Capacity} \\ \Phi & = & [x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 & \geq x_4 \otimes 8 + x_5 \otimes 14 + x_6 \otimes 2] \end{aligned}$$

- Assume x_i are Boolean random variables
- Then the sum expression $\alpha = x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9$ is a $\mathbb{N}$-valued random variable

Algebraic Expressions give rise to Random Variables

$$\begin{array}{rcccc} & & \text{Democratic Biomass Capacity} & \geq & \text{Republican Biomass Capacity} \\ \Phi & = & [x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 & \geq & x_4 \otimes 8 + x_5 \otimes 14 + x_6 \otimes 2] \end{array}$$

- Assume x_i are Boolean random variables
- Then the sum expression $\alpha = x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9$ is a $\mathbb{N}$-valued random variable
- Hence Φ is a $\mathbb{B}$-valued random variable

Algebraic Expressions give rise to Random Variables

$$\begin{array}{rcccc} & & \text{Democratic Biomass Capacity} & \geq & \text{Republican Biomass Capacity} \\ \Phi & = & [x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9 & \geq & x_4 \otimes 8 + x_5 \otimes 14 + x_6 \otimes 2] \end{array}$$

- Assume x_i are Boolean random variables
- Then the sum expression $\alpha = x_1 \otimes 17 + x_2 \otimes 5 + x_3 \otimes 9$ is a $\mathbb{N}$-valued random variable
- Hence Φ is a $\mathbb{B}$-valued random variable
- $P_\Phi[\top]$ is the probability that a random choice of possible values for the variables x_i satisfies the inequality
- In this example, $P_\Phi[\top]$ is the probability that Democrats are responsible for more biomass capacity than Republicans

Outline

Monoids, Semirings, Semimodule

What do we mean by $+$ in $\Phi_1 \otimes 17 + \Phi_2 \otimes 5$?

Well, it depends . . .

Monoids, Semirings, Semimodule

What do we mean by $+$ in $\Phi_1 \otimes 17 + \Phi_2 \otimes 5$?
Well, it depends . . .

Aggregation modelled by commutative monoids

- Carrier *M*, e.g. $\mathbb{N}$ or $\mathbb{R}$
- Binary operation $M \times M \to M$
- Neutral element $0 \in M$
- Examples for aggregation monoids:
  SUM $(\mathbb{N}, +, 0)$, MIN $(\mathbb{N}, \min, \infty)$, MAX $(\mathbb{N}, \max, -\infty)$,
  PROD, COUNT (special case of SUM)

Monoids, **Semirings**, Semimodule

What are Φ_1, Φ_2 in $\quad \Phi_1 \otimes 17 + \Phi_2 \otimes 5$?

Monoids, **Semirings**, Semimodule

What are Φ_1, Φ_2 in $\Phi_1 \otimes 17 + \Phi_2 \otimes 5$?

- Consider Query:

$$\text{AGG}_B\Big[(R \cup S) \bowtie_A T\Big]$$

R

A	Φ
1	x_1
2	x_2

S

A	Φ
1	y_1

T

A	B	Φ
1	17	z_1
2	5	z_2

Monoids, **Semirings**, Semimodule

What are Φ_1, Φ_2 in $\Phi_1 \otimes 17 + \Phi_2 \otimes 5$?

- Consider Query:

  $\text{AGG}_B\Big[(R \cup S) \bowtie_A T\Big]$

R

A	Φ
1	x_1
2	x_2

S

A	Φ
1	y_1

T

A	B	Φ
1	17	z_1
2	5	z_2

Tuples annotations modelled by semirings

- $(R \cup S) \bowtie_A T$ yields

$(R \cup S) \bowtie_A T$

A	B	Φ
1	17	$(x_1 + y_1) \cdot z_1$
2	5	$x_2 \cdot z_2$

- Aggregation on top of this table yields:
  $((x_1 + y_1) \cdot z_1) \otimes 17 + (x_2 \cdot z_2) \otimes 5$
  where the meaning of $+$ depends on the aggregation monoid

Monoids, Semirings, **Semimodule**

Semimodule

- Algebraic framework introduced by Amsterdamer et al. [2011]
- The algebraic structure combining semirings and monoids is called semimodule
- Generalisation of vector space. “Scalars”: tuple annotations, “Vectors”: aggregation values
- Semimodule expressions represent data values conditioned on tuple annotations

Semiring and semimodule expressions are random variables

- Semimodule: Random variable over aggregation domain
- Semiring expressions: ?
  - So far in probabilistic databases:
Boolean random variable
  - However: $\mathbb{B}$ is in general not large enough for aggregation; need larger semiring, for example natural numbers

Aggregation Needs Semirings Larger Than $\mathbb{B}$

ProducerEU

A	Φ
1	x_1
2	x_2

ProducerUS

A	Φ
1	y_1

Products

A	Price	Φ
1	17	z_1
2	5	z_2

- Query: $\mathrm{SUM}_{\mathrm{Price}}\left[(\mathrm{ProducerEU} \cup \mathrm{ProducerUS}) \bowtie_A \mathrm{Products}\right]$ asking for total price of products sold by all producers
- Resulting expression: $((x_1 + y_1) \cdot z_1) \otimes 17 + (x_2 \cdot z_2) \otimes 5$
- Valuation $\nu : x_1, x_2, y_1, z_1, z_2 \mapsto \top$ yields $\top \otimes 17 + \top \otimes 5 = 22$
  Arguably not the expected result

Aggregation Needs Semirings Larger Than $\mathbb{B}$

ProducerEU

A	Φ
1	x_1
2	x_2

ProducerUS

A	Φ
1	y_1

Products

A	Price	Φ
1	17	z_1
2	5	z_2

- Query: $\text{SUM}_{\text{Price}}\Big[(\text{ProducerEU} \cup \text{ProducerUS}) \bowtie_A \text{Products}\Big]$ asking for total price of products sold by all producers
- Resulting expression: $((x_1 + y_1) \cdot z_1) \otimes 17 + (x_2 \cdot z_2) \otimes 5$
- Valuation $\nu : x_1, x_2, y_1, z_1, z_2 \mapsto \top$ yields $\top \otimes 17 + \top \otimes 5 = 22$ Arguably not the expected result

- Boolean semiring is not large enough for SUM
- Better choice: Semiring $\mathbb{N}$. Identify $\bot \sim 0$, $\top \sim 1$.
- Valuation $\nu : x_1, x_2, y_1, z_1, z_2 \mapsto 1$ yields $((1 + 1) \cdot 1) \otimes 17 + (1 \cdot 1) \otimes 5 = 2 \otimes 17 + 1 \otimes 5 = 39.$

A More Formal View: **Expressions, Random Variables**

- The probability space *induced by* **X** has as samples the set of valuations from **X** to S,

$$\Omega = \{\nu : \mathbf{X} \to S\}$$

- Every expression $\Phi \in K$ is an S-valued random variable over Ω with probability distribution

$$P_{\Phi}[s] = P\big(\{\nu \in \Omega \mid \nu(\Phi){=}s\}\big) = \sum_{\substack{\nu\in\Omega: \\ \nu(\Phi)=s}} P(\nu)$$

for every $s \in S$

Outline

The pvc-tables Representation System

Ingredients for pvc-tables

- A set **X** of variable symbols
- Tuples contain constants or semimodule expressions over **X**
- Every tuple is annotated with a semiring expression over **X**

Queries

- Query Q maps pvc-table database D to pvc-table $Q(D)$
- Annotations are propagated via query operators
- Expressions concisely encode probability distributions of answers

Properties of pvc-tables

- Polynomial overhead (Amsterdamer et al. [2011]):
  $|Q(D)| \in \mathcal{O}\big(\mathrm{poly}(|D|)\big)$ (unlike pc-tables)
- Completeness: Every finite probability distribution over relations (with set or bag semantics) can be represented by pvc-tables

The pvc-tables Representation System

Semantics: Set vs Bag & Deterministic vs Probabilistic

Different choices for the semiring and the probability distributions of the annotation variables give rise to different database semantics.

Database Semantics		Semiring	Probability Distributions
Deterministic	Set	$\mathbb{B}$	$P_x[\top] = 1$ or $P_x[\bot] = 1$
Deterministic	Bag	$\mathbb{N}$	$\exists n \in \mathbb{N}: \; P_x[n] = 1$
Probabilistic	Set	$\mathbb{B}$	$P_x[\top], P_x[\bot] \in [0, 1]$
Probabilistic	Bag	$\mathbb{N}$	$\forall n \in \mathbb{N}: \; P_x[n] \in [0, 1]$

Outline

Query Evaluation in pvc-tables (1)

Step 1: Construction of Expressions

Alongside (standard) query evaluation, compute annotations.

- Project, Union, Cartesian Product: Construction of semiring expressions ($\cdot$ for joint, and $+$ for alternative use of data)
- Aggregation (with grouping): Construct semimodule expressions ($\sum_{\text{AGG}} \Phi \otimes v$)

R

A	B	Φ
a	1	x_1
a	2	x_2
b	3	x_3
b	4	x_4

$$\xrightarrow{\text{select AGG(B) from R group by A}}$$

pvc-table

A	AGG(B)	Φ
a	$x_1 \otimes 1 + x_2 \otimes 2$	$[x_1 + x_2 \neq 0]$
b	$x_3 \otimes 3 + x_4 \otimes 4$	$[x_3 + x_4 \neq 0]$

Query Evaluation in pvc-tables (1)

Step 1: Construction of Expressions

Alongside (standard) query evaluation, compute annotations.

- Project, Union, Cartesian Product: Construction of semiring expressions ($\cdot$ for joint, and $+$ for alternative use of data)
- Aggregation (with grouping): Construct semimodule expressions ($\sum_{\text{AGG}} \Phi \otimes v$)

R

A	B	Φ
a	1	x_1
a	2	x_2
b	3	x_3
b	4	x_4

$$\xrightarrow{\text{select AGG(B) from R group by A}}$$

pc-table

A	SUM(B)	Φ
a	0	$\bar{x}_1 \cdot \bar{x}_2$
a	1	$x_1 \cdot \bar{x}_2$
a	2	$\bar{x}_1 \cdot x_2$
a	3	$x_1 \cdot x_2$
b	0	$\bar{x}_3 \cdot \bar{x}_4$
...		

Exponential overhead!
Lechtenbörger et al. [2002]

Query Evaluation in pvc-tables (2)

Step 2: Probability Computation

Problem: Given a tuple, compute its probability distribution.

Idea: Tuple probability is equivalent to joint probability distribution of its semimodule expressions and annotation expression as obtained from evaluation step 1.

Approach: Compile expressions into a tractable form consisting of *independent* and *mutually exclusive* sub-expressions.

Compilation: Independent Decomposition

Consider semiring expression $\Phi = x + y$. Since x, y are independent random variables, the probability distribution of Φ is given by the convolution of x and y.

If x, y are in $\mathbb{N}$: $$P_{x+y}[n] = \sum_{\substack{i,j \in \mathbb{N} \\ i+j=n}} P_x[i] P_y[j]$$

Compilation: Independent Decomposition

Consider semiring expression $\Phi = x + y$. Since x, y are independent random variables, the probability distribution of Φ is given by the convolution of x and y.

If x, y are in $\mathbb{N}$: $$P_{x+y}[n] = \sum_{\substack{i,j \in \mathbb{N} \\ i+j=n}} P_x[i] P_y[j]$$

If x, y are Boolean: $$P_{x+y}[\bot] = \sum_{\substack{a,b \in \{\bot,\top\} \\ a \vee b = \bot}} P_x[a] P_y[b]$$

$$P_{x+y}[\top] = \sum_{\substack{a,b \in \{\bot,\top\} \\ a \vee b = \top}} P_x[a] P_y[b]$$

Compilation: Independent Decomposition

Consider semiring expression $\Phi = x + y$. Since x, y are independent random variables, the probability distribution of Φ is given by the convolution of x and y.

If x, y are in $\mathbb{N}$:
$$P_{x+y}[n] = \sum_{\substack{i,j \in \mathbb{N} \\ i+j=n}} P_x[i] P_y[j]$$

If x, y are Boolean:
$$P_{x+y}[\bot] = \sum_{\substack{a,b \in \{\bot,\top\} \\ a \vee b = \bot}} P_x[a] P_y[b] = P_x[\bot] P_y[\bot]$$
$$\begin{aligned} P_{x+y}[\top] &= \sum_{\substack{a,b \in \{\bot,\top\} \\ a \vee b = \top}} P_x[a] P_y[b] \\ &= P_x[\top]P_y[\top] + P_x[\bot]P_y[\top] + P_x[\top]P_y[\bot] \\ &= 1 - P_x[\bot]P_y[\bot] \end{aligned}$$

Compilation: Independent Decomposition

The applicability of convolution is not limited to “sums”; convolution is equally well defined for other binary operations:

Convolution for algebraic operations

- Semiring expressions: $\Phi \cdot \Psi$, $\Phi + \Psi$
- Semimodule expressions: $\alpha + \beta$
- Mixed semiring and semimodule expressions: $\Phi \otimes \alpha$
- Convolution is also applicable to comparisons of expressions, such as $\alpha \leq \beta$

Compilation: Mutually Exclusive Expressions

What if there are no independent sub-expressions?
Example: $\alpha = a(b + c) \otimes 10 + c \otimes 20$

Idea: Instantiate one of the variables to create mutually exclusive sub-expressions.

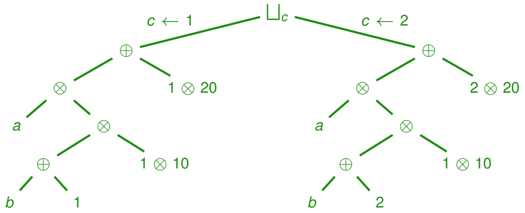
$$\begin{aligned} P(\alpha) = {} & P_c[1] \cdot P\big(a(b+1) \otimes 10 + 1 \otimes 20\big) + \\ & P_c[2] \cdot P\big(a(b+2) \otimes 10 + 2 \otimes 20\big) + \\ & P_c[3] \cdot P\big(a(b+3) \otimes 10 + 3 \otimes 20\big) + \\ & \dots \end{aligned}$$

Need to consider all possible values of c with non-zero probability.
In particular: For Boolean variables, the above construction yields Shannon's expansion.

Decomposition Trees (d-trees)

Decomposition gives rise to a tree whose nodes explain the decomposition steps taken. For example, $\bigsqcup$ for mutex decomposition, $\oplus$ for convolution w.r.t. $+$, $\otimes$ for convolution w.r.t. $\otimes$, etc.

Example: $\alpha = a(b + c) \otimes 10 + c \otimes 20$



Tractable Probability Computation for d-trees

The probability distribution P_d of a d-tree d whose nodes have probability distributions $p_1, \ldots, p_n$ can be computed in time $\mathcal{O}(\prod |p_i|)$.

Specific polynomial time cases

- For MIN and MAX monoids combined with any semiring
- For SUM monoid: If monoid values and size of probability distributions of semiring expressions are bounded by constants
    - This subsumes COUNT aggregation

Further Applications of d-trees

- Approximate probability computation by partial expansion of d-tree (Olteanu et al. [2010], Fink et al. [2011])
- Sensitivity analysis and explanation of query results (Kanagal et al. [2011])
- Conditioning probabilistic databases (Koch and Olteanu [2008])

Tractable Queries via d-trees

Tractability for query evaluation on probabilistic databases is considered with respect to data complexity:

For which class of queries can probability distributions of query answers be computed in *polynomial-time data complexity* for any tuple-independent database?

Tractable Queries via d-trees

Tractability for query evaluation on probabilistic databases is considered with respect to data complexity:

For which class of queries can probability distributions of query answers be computed in *polynomial-time data complexity* for any tuple-independent database?

- Syntactic characterisation of tractable queries with aggregates
  - There are known classes of tractable non-aggregate queries with polynomial-time d-tree compilation, e.g. hierarchical queries
  - Extend these classes by adding nested aggregation without breaking the tractable (e.g. hierarchical) property

Tractable Queries via d-trees

Example 1

```
select R.A from R where R.B = (select MIN(S.B) from S
                                      where S.C = R.C)
```

Tractable sub-queries without aggregation:

```
select S.B from S where S.C = R.C
```

Tractable Queries via d-trees

Example 2

select 1 where

$$\Big(\text{select MIN(R.A) from R}\Big) <= \Big(\text{select COUNT(*) from S,T where S.A=T.A}\Big)$$

Tractable sub-queries without aggregation:

select 1 where (select R.A from R)
select 1 from S,T where S.A=T.A

select 1 where (select R.A from R) <= (select 1 from S,T where S.A=T.A)

Outline

Performance Analysis

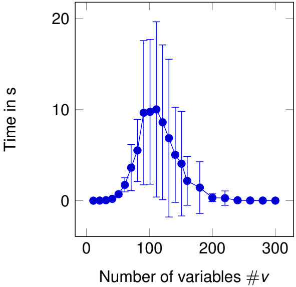


Figure: Varying the number of variables for a randomly generated semimodule expression (L=90, $\#cl$=2, $\#l$=2, $maxv$=5, c=3, $\#runs$=40, AGGL=MIN)

$$\left[\sum_{\text{AGGL}}^{L} \Phi_i \otimes v_i \;=\; c \right]$$

Performance Analysis

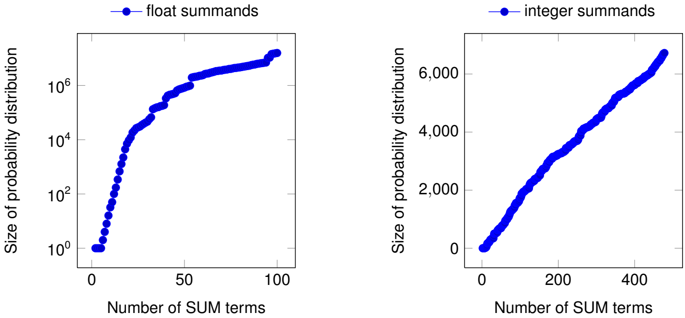


Figure: Size of the probability distributions for SUM semimodule expressions of varying size. When summing float numbers from a fixed range, the size of the probability distribution grows potentially exponentially in the number of terms, while summing integers from a fixed range it grows linearly.

Performance Analysis

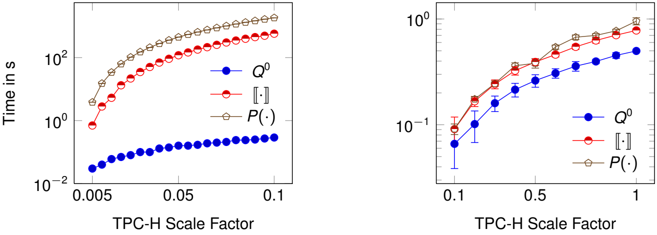
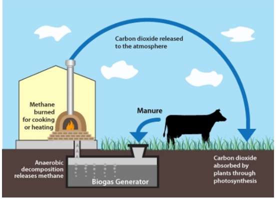


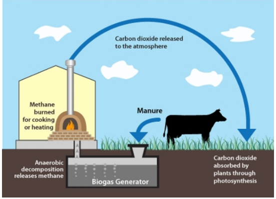
Figure: TPC-H Queries Q1 (modified) and Q2. For each query, the graphs compare the execution times (1) on a deterministic database (Q^0) without expression or probability computation, (2) of the computation of the expressions ($[\![\cdot]\!]$), and (3) of probability computation for the result tuples ($P(\cdot)$).


Carbon dioxide released to the atmosphere
Methane burned for cooking or heating
Manure
Anaerobic decomposition releases methane
Biogas Generator
Carbon dioxide absorbed by plants through photosynthesis

?

Who is responsible for a larger capacity of biogas plants, Democrats or Republicans?

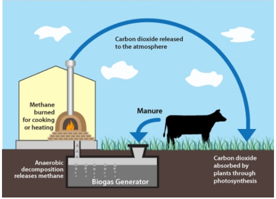
?


Carbon dioxide released to the atmosphere
Methane burned for cooking or heating
Manure
Anaerobic decomposition releases methane
Biogas Generator
Carbon dioxide absorbed by plants through photosynthesis

?

Who is responsible for a larger capacity of biogas plants, Democrats or Republicans?

?



55%

Who is responsible for a larger capacity of biogas plants, Democrats or Republicans?

45%

End. ?

Definitions

Monoid

A monoid is a set M with an operation $+ : M \times M \to M$ and a neutral element $0 \in M$ that satisfy the following axioms for all $m_1, m_2, m_3 \in M$:

$$(m_1 + m_2) + m_3 = m_1 + (m_2 + m_3)$$
$$0 + m_1 = m_1 + 0 = m_1$$

A monoid is commutative if $m_1 + m_2 = m_2 + m_1$

Semiring

A commutative semiring is a set S together with operations $+, \cdot : S \times S \to S$ and neutral elements $0, 1 \in S$ such that $(S, +, 0)$ and $(S, \cdot, 1)$ are commutative monoids and the following holds for all $s_1, s_2, s_3 \in S$:

$$s_1 \cdot (s_2 + s_3) = (s_1 \cdot s_2) + (s_1 \cdot s_3)$$
$$(s_1 + s_2) \cdot s_3 = (s_1 \cdot s_3) + (s_2 \cdot s_3)$$
$$0 \cdot s_1 = s_1 \cdot 0 = 0$$

Definitions

Semimodule

Let $(S, +_S, 0_S, \cdot_S, 1_S)$ be a commutative semiring. As *S*-semimodule M consists of a commutative monoid $(M, +_M, 0_M)$ and a binary operation $\otimes : S \times M \to M$ such that for all $s_1, s_2 \in S$ and $m_1, m_2 \in M$ we have

$$\begin{aligned}
s_1 \otimes (m_1 +_M m_2) &= s_1 \otimes m_1 +_M s_1 \otimes m_2 \\
(s_1 +_S s_2) \otimes m_1 &= s_1 \otimes m_1 +_M s_2 \otimes m_1 \\
(s_1 \cdot_S s_2) \otimes m_1 &= s_1 \otimes (s_2 \otimes m_1) \\
s_1 \otimes 0_M = 0_K \otimes m_1 &= 0_M \\
1_S \otimes m_1 &= m_1
\end{aligned}$$

Further Experiments

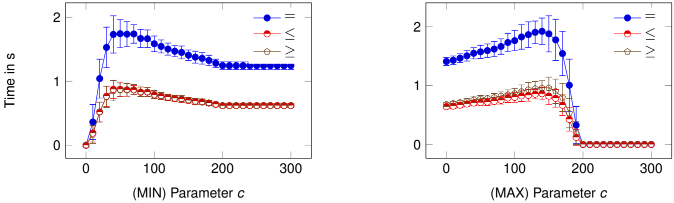


Figure: Experiment A: Varying the constant *c* for different aggregation monoids and comparison operators θ. $\#v=25$, $L=200$, $R=0$, $\#cl=3$, $\#l=3$, $maxv=200$.

Further Experiments

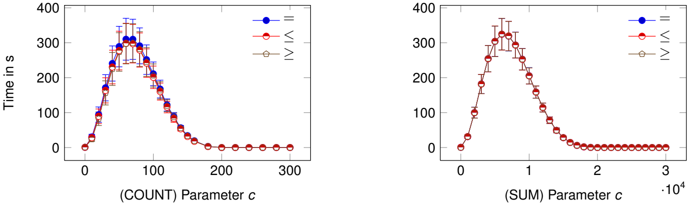


Figure: Experiment A: Varying the constant *c* for different aggregation monoids and comparison operators θ. $\#v$=25, *L*=200, *R*=0, $\#cl$=3, $\#l$=3, *maxv*=200.

References