

Default-all is dangerous!

Wolfgang Gatterbauer
Alexandra Meliou
Dan Suciu

3rd USENIX Workshop on the Theory and Praxis of Provenance (Tapp'11)

Overview Provenance Definitions

	Why?	Where?
<i>Naive</i>	Witness	"SQL interpretation"
<i>Provenance definition</i>	<p><u>Why-provenance = witness basis (α_w)</u></p> <p>Buneman et al. [ICDT'01]</p>	<p><u>Where-provenance = propagation (α_p)</u></p> <p>Buneman et al. [PODS'02]</p>
<i>QRI definition (Query-Rewrite-Insensitive)</i>	<p><u>Minimal witness basis (α_w^m)</u></p> <p>Buneman et al. [ICDT'01]</p>	<p><u>Default-all propagation (α_p^d)</u></p> <p>Bhagwat et al. [VLDB'04]</p>

We do not discuss here whether QRI is desirable (see also Glavic, Miller [Tapp'11]), but merely point out that, if aiming for QRI, care has to be taken about the ramifications of the proposed semantics.

Has problems if one interprets annotations on attribute values

Independent work presented at this WS

Minimal propagation (α_p^m)
Proposed in this paper!

Overview Provenance Definitions

		Why?	Where?																																																																												
<i>Naive</i>		Witness	"SQL interpretation"																																																																												
<i>Provenance definition</i>		<p>Why-provenance = witness basis (α_w)</p> <p>Buneman et al. [ICDT'01]</p>	<p>Where-provenance = propagation (α_p)</p> <p>Buneman et al. [PODS'02]</p>																																																																												
	<p>Glavic, Miller [Tapp'11]</p> <table border="1"> <thead> <tr> <th>Semantics</th> <th>Sound</th> <th>Complete</th> <th>Responsible</th> <th>Insensitive (set)</th> <th>Insensitive (bag)</th> <th>Stable</th> </tr> </thead> <tbody> <tr> <td>Why</td> <td>Wit</td> <td>- X</td> <td>-</td> <td>X X</td> <td>X X</td> <td>X X</td> </tr> <tr> <td></td> <td>Why</td> <td>- X</td> <td>-</td> <td>-</td> <td>X X</td> <td>X X</td> </tr> <tr> <td></td> <td>IWhy</td> <td>- X</td> <td>X X</td> <td>X X</td> <td>X X</td> <td>X X</td> </tr> <tr> <td>Where</td> <td>Where</td> <td>- -</td> <td>- -</td> <td>-</td> <td>?</td> <td>X</td> </tr> <tr> <td></td> <td>IWhere</td> <td>- -</td> <td>- -</td> <td>X</td> <td>X</td> <td>-</td> </tr> <tr> <td>How</td> <td></td> <td>- X</td> <td>- -</td> <td>-</td> <td>X X</td> <td>X X</td> </tr> <tr> <td>Lineage-based</td> <td>Lineage</td> <td>X X</td> <td>- -</td> <td>- -</td> <td>-</td> <td>X</td> </tr> <tr> <td></td> <td>PI-CS</td> <td>X X</td> <td>- -</td> <td>- -</td> <td>-</td> <td>X</td> </tr> <tr> <td></td> <td>C-CS</td> <td>X</td> <td>- -</td> <td>- -</td> <td>-</td> <td>X</td> </tr> <tr> <td>Causality</td> <td></td> <td>- X</td> <td>X X</td> <td>X X</td> <td>X X</td> <td>X X</td> </tr> </tbody> </table>	Semantics	Sound	Complete	Responsible	Insensitive (set)	Insensitive (bag)	Stable	Why	Wit	- X	-	X X	X X	X X		Why	- X	-	-	X X	X X		IWhy	- X	X X	X X	X X	X X	Where	Where	- -	- -	-	?	X		IWhere	- -	- -	X	X	-	How		- X	- -	-	X X	X X	Lineage-based	Lineage	X X	- -	- -	-	X		PI-CS	X X	- -	- -	-	X		C-CS	X	- -	- -	-	X	Causality		- X	X X	X X	X X	X X	<p>Default-all propagation (α_p^d)</p> <p>Bhagwat et al. [VLDB'04]</p> <p>Has problems if one interprets annotations on attribute values</p> <p>Note that Minimal propagation is "stable", in contrast to Default-all</p> <p>Minimal propagation (α_p^m) Proposed in this paper!</p>
Semantics	Sound	Complete	Responsible	Insensitive (set)	Insensitive (bag)	Stable																																																																									
Why	Wit	- X	-	X X	X X	X X																																																																									
	Why	- X	-	-	X X	X X																																																																									
	IWhy	- X	X X	X X	X X	X X																																																																									
Where	Where	- -	- -	-	?	X																																																																									
	IWhere	- -	- -	X	X	-																																																																									
How		- X	- -	-	X X	X X																																																																									
Lineage-based	Lineage	X X	- -	- -	-	X																																																																									
	PI-CS	X X	- -	- -	-	X																																																																									
	C-CS	X	- -	- -	-	X																																																																									
Causality		- X	X X	X X	X X	X X																																																																									

Example 1: Query-Rewrite-Insensitivity (QRI)

Why

Input

R

	A	B
t ₁	1	2
t ₂	1	3
t ₃	2	2

Why-provenance = witness basis (α_w)

Query 1

$Q_1(x,y):-R(x,y)$

	A	B
	1	2
	1	3
	2	2

Query 2 \equiv *Query 1*

$Q_2(x,y):-R(x,y), R(_,y)$

	A	B
	1	2
	1	3
	2	2

Minimal witness basis (α_w^m)

Lineage (α_l)

{t₁, t₃}
{t₂}
{t₁, t₃}

Where

Input

R^a

	A	B
	1 ^a	2 ^b
	1 ^c	3 ^d
	2 ^e	2 ^f

Where-provenance = propagation (α_p)

Query 1

$Q_1(x,y):-R^a(x,y)$

	A	B
	1 ^a	2 ^b
	1 ^c	3 ^d
	2 ^e	2 ^f

Minimal propagation (α_p^m)

Default-all propagation (α_p^d)

Query 2 \equiv *Query 1*

$Q_2(x,y):-R^a(x,y), R^a(_,y)$

	A	B
	1 ^a	2 ^{b,f}
	1 ^c	3 ^d
	2 ^e	2 ^{b,f}

	A	B
	1 ^{a,c}	2 ^{b,f}
	1 ^{a,c}	3 ^d
	2 ^e	2 ^{b,f}

	A	B
	1 ^a	2 ^b
	1 ^c	3 ^d
	2 ^e	2 ^f

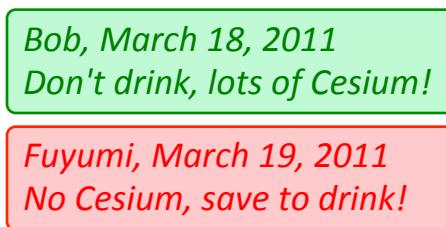
Real example: Why Default-all is dangerous

Hanako queries a community DB for contents of LF-milk*:

Community Database

R^a

Food	Content
LF Milk	Cesium-137 ^b
LF Milk	Calcium ^d
SC Water	Cesium-137 ^f



Hanako's query

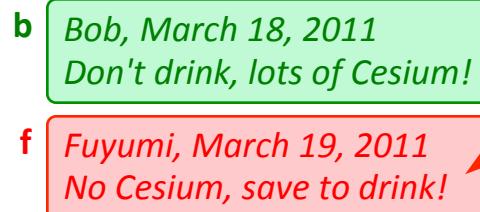
$Q(y) :- R^a(\text{`LF Milk'}, y)$

Content
Cesium-137 ^{???}
Calcium ^d

Default-all propagation makes her drink the milk:

Default-all propagation ($\alpha_p^{d\prime}$)

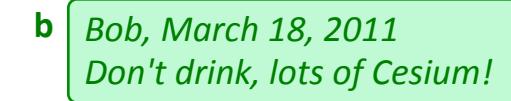
Content
Cesium-137 ^{bf}
Calcium ^d



"semantically irrelevant information": annotations leak over from SC Water tuple to LF Milk

Minimal propagation (α_p^m)

Content
Cesium-137 ^b
Calcium ^d



"all relevant and only relevant"

* Note the one-to-one correspondence of this example with example 1

Definition Minimal propagation (α_p^m)

$$\alpha_p^m(t, A, Q) := \bigcup_{\substack{t' \in \bigcup \alpha_w^m(t, Q) \\ A' \in \text{attributes of } t' \text{ propagating to cell}(t, A)}} \alpha_p(t', A')$$

\bigcup transforms 'sets of sets' into 'sets',
hence something like QRI lineage

Intuition:

Return the intersection between:

- query-specific where-provenance (α_p)
- and QRI minimal witness basis (α_w^m)

"all relevant ... and only relevant"

Example 1

Input

R^a

	A	B
t ₁	1 ^a	2 ^b
t ₂	1 ^c	3 ^d
t ₃	2 ^e	2 ^f

Where provenance (α_p)

Query 2

$Q_2(x, y) :- R^a(x, y), R^a(_, y)$

A	B
1 ^a	2 ^{b,f}
1 ^c	3 ^d
2 ^e	2 ^{b,f}

$\{t_1\}$

$\{t_2\}$

$\bigcup \alpha_w^m$

Minimal propagation (α_p^m)

A	B
t ₄	1 ^a
t ₅	1 ^c
t ₆	2 ^e

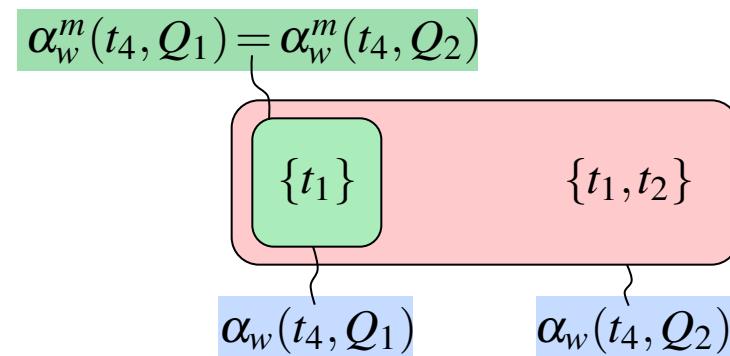
$$\begin{aligned} \alpha_p^m(t_4, B, Q_2) &= \bigcup_{t' \in \{t_1\}, A'} \alpha_p(t', A') \\ &= \alpha_p(t_1, B) = \{b\} \end{aligned}$$

Example 1: Illustration of "minimal" versus "all"

Why-provenance

Why-provenance (α_w)

Minimal witness basis (α_w^m)

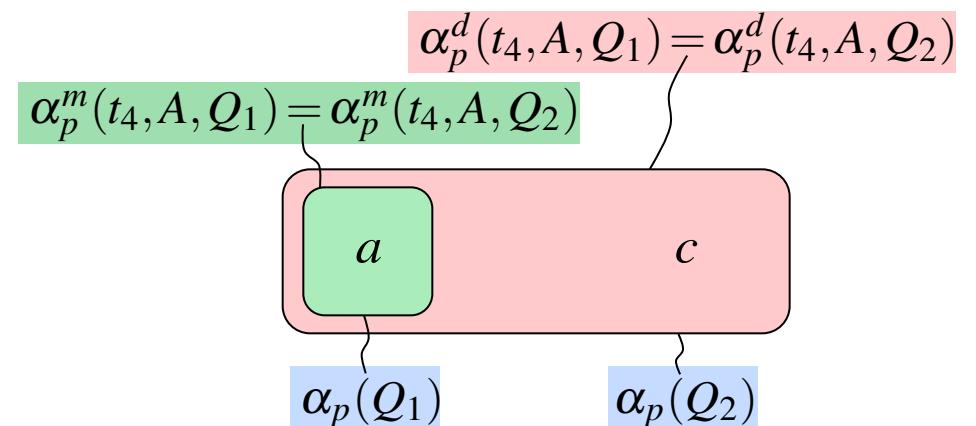


Where-provenance

Where-provenance (α_p)

Default-all propagation (α_p^d)

Minimal propagation (α_p^m)



Interpretation of Annotations 1: Attribute Value*



	Item Name	Description	Population	
athens heraklion chania				
<input checked="" type="checkbox"/>	athens	PIRAEUS (Athens) - HERAKLION (Crete) - PIRAEUS (Athens) . PIRAEUS (Athens) - CHANIA (Crete) - PIRAEUS (Athens)	4 possible values	
<input checked="" type="checkbox"/>	heraklion	Heraklion or Iraklion is the largest city and capital of Crete. It is also the 4th largest city in Greece. Heraklion is the capital of	1 possible value	
<input checked="" type="checkbox"/>	kania	Chania confusingly is sometimes written Hania though it can also be written Khania, Cania, Canea and Kania and in Greek is Χανιά	1 possible value	
<input checked="" type="checkbox"/>	Crete	A superb way of enjoying the journey to Crete is to fly to Athens and take the ferry from Piraeus (Piraeus), the port serving Athens	623,666	
<input checked="" type="checkbox"/>	Mykonos	Heraklion and Chania are international airports, Sitia airport is currently receiving domestic flights only charter flights are expected to	9,320	
<input checked="" type="checkbox"/>	Istanbul	14 Days - Depart USA, stops include, Istanbul , Mount Athos, Skithos, Samos, Kusadasi, Delos,	8,260,000	

* Interpretation of annotations on entity attribute values favored by us and underlying our model

Interpretation of Annotations 1: Attribute Value*



athens heraklion chania			
Item Name	Description	Population	
athens	PIRAEUS (Athens) - HERAKLION (Crete) - PIRAEUS (Athens). PIRAEUS (Athens) - CHANIA (Crete). DIPACUS (Athens)		<p>Annotations on values of an attribute (here "population") for a particular entity (here "Athens")</p>
heraklion	Heraklion or Iraklion is the large city and capital of Crete. It is also the 4th largest city in Greece. Heraklion is the capital of		<p>Possible values</p> <ul style="list-style-type: none"><input checked="" type="radio"/> 750000 Low confidence Greece. LOCATION. Official Website: http://www.cityofathens.gr/. Population: 750000. Population of Athens metropolitan area, 3.7 million www.nndb.com - all 2 sources »<input type="radio"/> 22936, 24234 Low confidence Population for Athens www.freebase.com<input type="radio"/> 1,102 Low confidence pop. for Athens www.citytowninfo.com<input type="radio"/> 18,967 Low confidence pop. for Athens www.citytowninfo.com - all 2 sources »
kania	Chania confusingly is sometimes written Hania though it can also be written Khania, Cania, Canea and Kanis and in Greek is Χανιά.		
Crete	A superb way of enjoying the journey to Crete is to fly to Athens and take the ferry from Piraeus (Pireas), the port serving Athens.		
Mykonos	Heraklion and Chania are international airports, Sitia airport currently receiving domestic flights only (charter flights are expected).		
Istanbul	14 Days - Depart USA, stops include, Istanbul, Mount Athos, Skithos, Samos, Kusadasi, Delos,		

Argument: Interpreting cell annotations as relevant to the tuple (entity) adds something that is not trivially modeled with normalized tables.

* Interpretation of annotations on entity attribute values favored by us and underlying our model

Interpretation of Annotations 2: Domain Value*

Domain value annotations*

Input R^a:

A	B	
1 ^a	2 ^b	<i>Bob, March 18, 2011 This number is a prime number.</i>
1 ^c	3 ^d	
2 ^e	2 ^f	<i>Fuyumi, March 19, 2011 Two is not a prime number because it is even.</i>

Input S^a:

...	Date	
...	Dec 25	<i>This is a holiday.</i>
...	...	
...	Dec 25	<i>This is a holiday too !!!</i>

Argument for default-all: If annotations are on domain values, then retrieving all annotations are relevant.

Alternative representation

Annotation table S^a:

B	annotation
2	<i>b: Bob, March 18, 2011 This number is a prime number.</i>
2	<i>f: Fuyumi, March 19, 2011 Two is not a prime number because it is even</i>

Annotation table S^a:

Date	annotation
Dec 25	<i>This is a holiday.</i>

Counter-Argument: But then these annotations can be modeled in a separate table as normalized tables.

* Alternative interpretation suggested by Wang-Chiew Tan (example created after conversation at Sigmod 2011)

Backup: Detailed Example 2

R^a	A	B
t_1	1^a	2^b
t_2	1^c	3^d
t_3	2^e	2^f
t_4	2^g	4^h

$Q_5(x,y) :- R^a(x,y), R^a(y,_), R^a(x,_)$

	<i>A</i>	<i>B</i>
<i>t</i> ₅	1 ^{a,c}	2 ^{b,e,g}
<i>t</i> ₆	2 ^{e,g}	2 ^{e,f,g}

$\{\{t_1, t_3\}, \{t_1, t_2, t_3\}, \{t_1, t_4\}, \{t_1, t_2, t_4\}\}$
 $\{\{t_3\}, \{t_3, t_4\}\}$

$$\{\{t_1, t_3\}, \{t_1, t_4\}\}$$

$$\{t_1, t_3, t_4\}$$

Where-provenance (α_p)

Why-provenance (α_w)

Minimal witness basis (α_w^m)

Default-all propagation (α_p^d)

<i>A</i>	<i>B</i>
1 a, c	2 b, e, f, g
2 e, g	2 b, e, f

$$\alpha_p^d(t_4 B, Q_5) = \alpha_p(t_4 B, Q_6) \quad \text{with}$$

$Q_6(x,y):-R^a(x,y), R^a(y,_), R^a(x,_), S^a(_,y)$

Minimal propagation (α_p^m)

	A	B
t ₄	1 ^a	2 ^{b,e,g}
t ₅	2 ^e	2 ^{e,f}

$$\begin{aligned}\alpha_p^m(t_4, A, Q_5) &= \bigcup_{t' \in \{t_1, t_3, t_4\}, A'} \alpha_p(t', A') \\ &= \alpha_p(t_1, A) = \{a\}\end{aligned}$$

$$\alpha_p^m(t_5, B, Q_5) = \bigcup_{t' \in \{t_3\}, A'} \alpha_p(t', A')$$

—————> $= \alpha_p(t_3, B) \cup \alpha_p(t_3, A) = \{e, f\}$

Note minimal propagation is not equivalent to just evaluating the where-provenance for the query:

$Q_7(x,y) :- R^a(x,y), R^a(y,_)$. E.g. $\alpha_p(t_5, B, Q_7) = \{e, f, g\}$