Version June 20, 2011

Default-all is dangerous!

Wolfgang Gatterbauer
Alexandra Meliou
Dan Suciu

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

Database group
University of Washington

http://db.cs.washington.edu/causality/

Overview Provenance Definitions

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

Has problems if one interprets annotations on attribute values

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

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.

Independent work presented at this WS

Overview Provenance Definitions

	Why?	Where?
Naive	Witness	"SQL interpretation"
Provenance definition	Why-provenance = witness basis (α_w) Buneman et al. [ICDT'01]	Where-provenance = propagation (α_p) Buneman et al. [PODS'02]

(α_w^m)

DT'01]

Default-all propagation (α_p^d)

Bhagwat et al. [VLDB'04]

Glavic, Miller [Tapp'11]

Semantics		Sound	Complete	Responsible	Insensitive (set)	Insensitive (bag)	Stable
Why	Wit	-	X	-	X	X	X
	Why	-	X	-	-	X	X
	IWhy	-	X	X	X	X	X
Where	Where	-	-	-	-	?	X
	IWhere	-	-	-	X	X	-
How		-	X	-	-	X	X
Lineage-based	Lineage	X	X	-	-	-	X
	PI-CS	X	X	-	-	-	X
	C-CS	X	-	-	-	-	X
Causality		-	X	X	X	X	X

Has problems if one interprets annotations on attribute values

Minimal propagation (α_p^m)

Proposed in this paper!

Note that Minimal propagation is "stable", in contrast to Default-all

Example 1: Query-Rewrite-Insensitivity (QRI)

Why

Why-provenance = witness basis (α_w)

Minimal witness basis (α_w^m)

Lineage (α_l)

Input
R

	A	B
t_1	1	2
t_2	1	3
t_3	2	2

Query 1
$Q_1(x,y)$:-$R(x,y)$

A	B	
1	2	$\{\{t_1\}\}$
1	3	$\{\{t_2\}\}$
2	2	$\{\{t_3\}\}$

Query 2 ≡ Query 1
$Q_2(x,y)$:-$R(x,y),R(\_,y)$

A	B			
1	2	$\{\{t_1\},\{t_1,t_3\}\}$	$\{\{t_1\}\}$	$\{t_1,t_3\}$
1	3	$\{\{t_2\}\}$	$\{\{t_2\}\}$	$\{t_2\}$
2	2	$\{\{t_3\},\{t_1,t_3\}\}$	$\{\{t_3\}\}$	$\{t_1,t_3\}$

Where

Where-provenance = propagation (α_p)

Minimal propagation (α_p^m)

Default-all propagation (α_p^d)

Input
R^a

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

Query 1
$Q_1(x,y)$:-$R^a(x,y)$

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

Query 2 ≡ Query 1
$Q_2(x,y)$:-$R^a(x,y),R^a(\_,y)$

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

Example adapted from Cheney et al. [Foundations and Trends in DBs'09]

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]

b *Bob, March 18, 2011*
Don't drink, lots of Cesium!

f *Fuyumi, March 19, 2011*
No Cesium, save to drink!

Hanako's query
Q (y):-R^a('LF Milk',y)

Content
Cesium-137[???]
Calcium[d]

Default-all propagation makes her drink the milk:

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

Content
Cesium-137[bf]
Calcium[d]

b *Bob, March 18, 2011*
Don't drink, lots of Cesium!

f *Fuyumi, March 19, 2011*
No Cesium, save to drink!

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

Minimal propagation ($\alpha_p{}^m$)

Content
Cesium-137[b]
Calcium[d]

b *Bob, March 18, 2011*
Don't drink, lots of Cesium!

"all relevant and only relevant"

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

Definition Minimal propagation (α_p^m)

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

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

Intuition:
Return the intersection between:
- query-specific where-provenanc (α_p)
- and QRI minimal witness basis (α_w^m)

"all relevant ... and only relevant"

Example 1

Where provenance (α_p)

Minimal propagation (α_p^m)

Input

R^a

	A	B
t_1	1^a	2^b
t_2	1^c	3^d
t_3	2^e	2^f

Query 2

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

A	B		
1^a	$2^{b,f}$	$\{\{t_1\}\}$	$\{t_1\}$
1^c	3^d	$\{\{t_2\}\}$	$\{t_2\}$
2^e	$2^{b,f}$	$\{\{t_3\}\}$	$\{t_3\}$

$\Cup \alpha_w^m$

Minimal witness basis (α_w^m)

	A	B
t_4	1^a	2^b
t_5	1^c	3^d
t_6	2^e	2^f

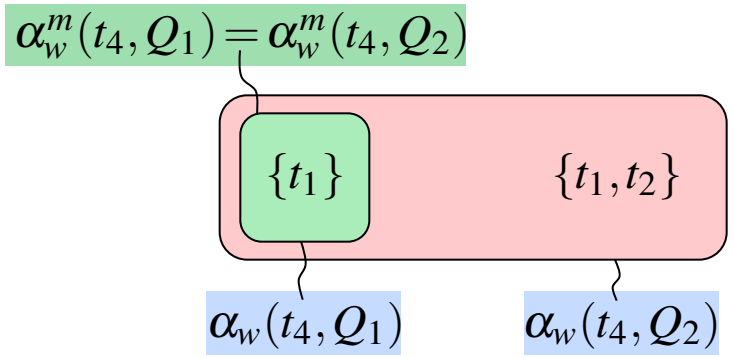
$$\alpha_p^m(t_4,B,Q_2) = \bigcup_{t' \in \{t_1\},A'} \alpha_p(t',A') = \alpha_p(t_1,B) = \{b\}$$

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

Why-provenance

Why-provenance (α_w)

Minimal witness basis ($\alpha_w{}^m$)

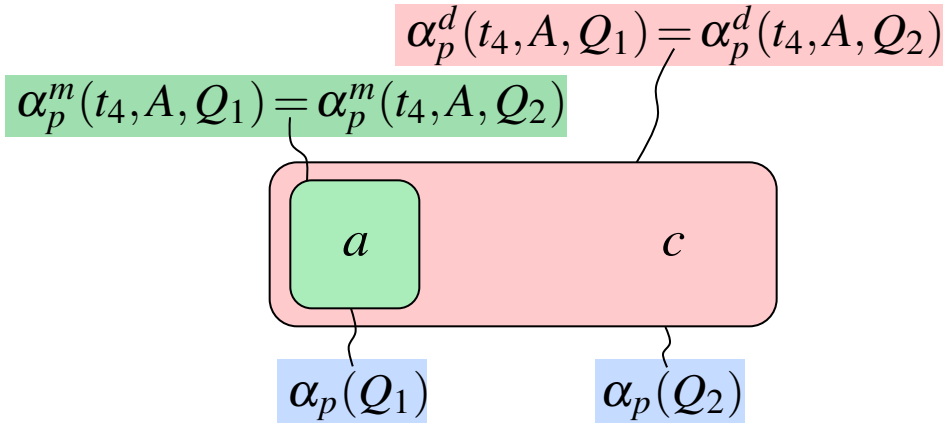


Where-provenance

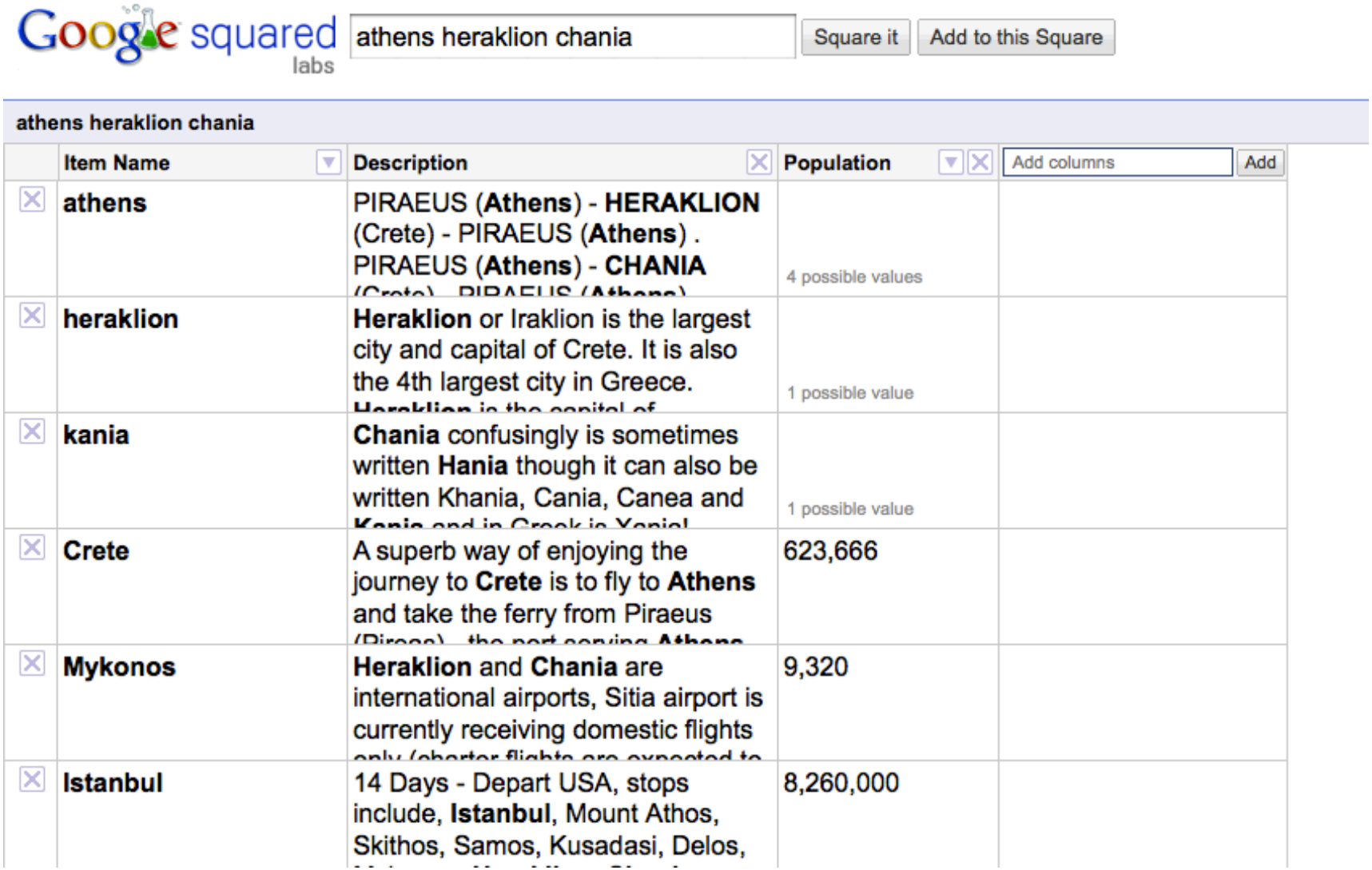
Where-provenance (α_p)

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

Minimal propagation ($\alpha_p{}^m$)

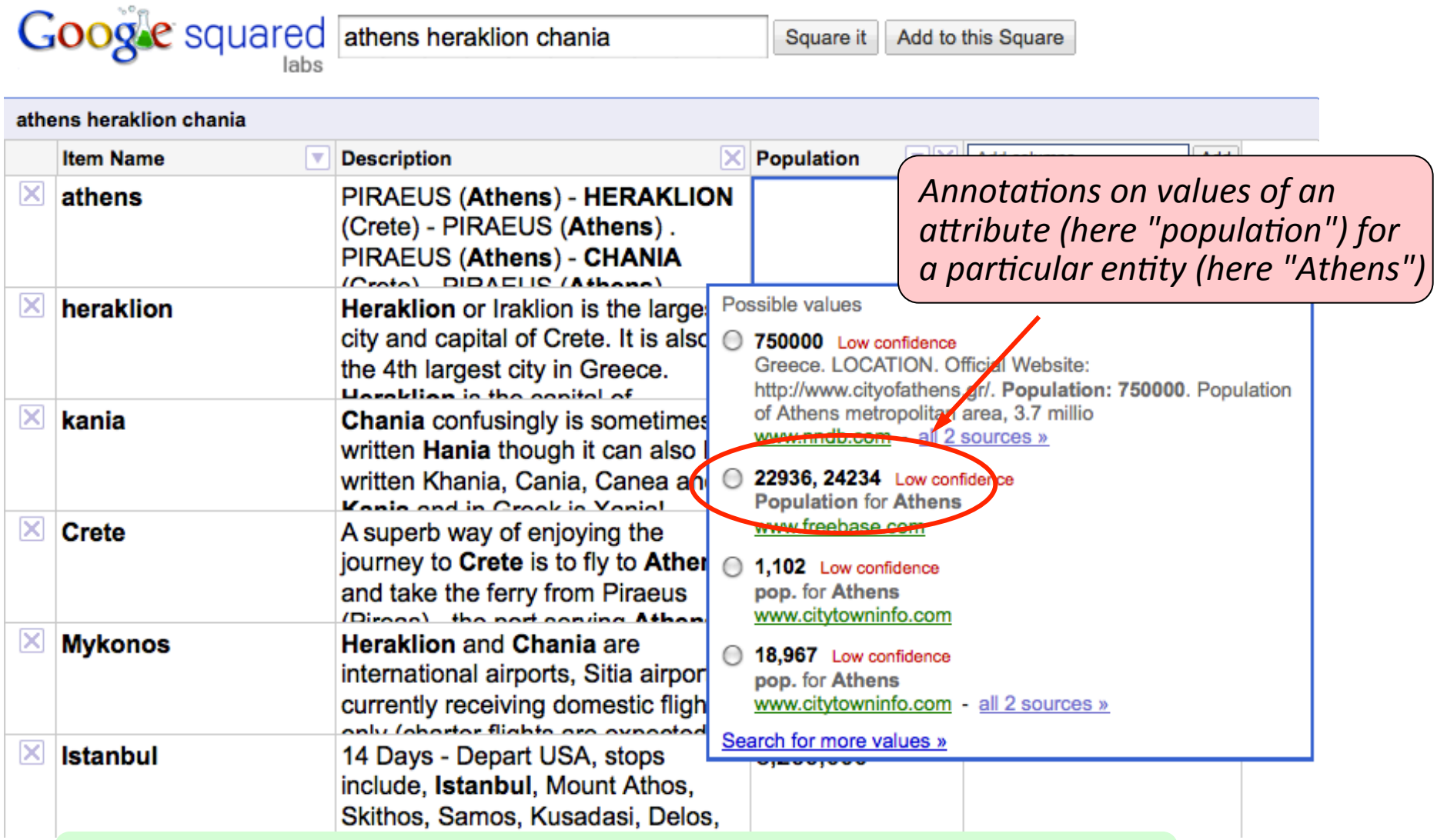


Interpretation of Annotations 1: Attribute Value*



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

Interpretation of Annotations 1: Attribute Value*



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
1^c	3^d
2^e	2^f

b: *Bob, March 18, 2011*
This number is a prime number.

f: *Fuyumi, March 19, 2011*
Two is not a prime number because it is even.

Input S^a:

...	Date
...	Dec 25
...	...
...	Dec 25

b: *This is a holiday.*

f: *This is a holiday too !!!*

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	b: *Bob, March 18, 2011* *This number is a prime number.*
2	f: *Fuyumi, March 19, 2011* *Two is not a prime number because it is even*

Annotation table S^a:

Date	annotation
Dec 25	This is a holiday.

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

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

	A	B
t_5	$1^{a,c}$	$2^{b,e,g}$
t_6	$2^{e,g}$	$2^{e,f,g}$

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

Why-provenance (α_w)

Where-provenance (α_p)

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

Minimal witness basis (α_w^m)

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

$\cup\alpha_w^m$ (~QRI lineage)

Default-all propagation (α_p^d)

A	B
$1^{a,c}$	$2^{b,e,f,g}$
$2^{e,g}$	$2^{b,e,f}$

$\alpha_p^d(t_4,B,Q_5) = \alpha_p(t_4,B,Q_6)$ 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	$2^{e,f}$

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

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

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. $\alpha_p(t_5,B,Q_7) = \{e,f,g\}$