Undecidability of determinacy of conjunctive queries

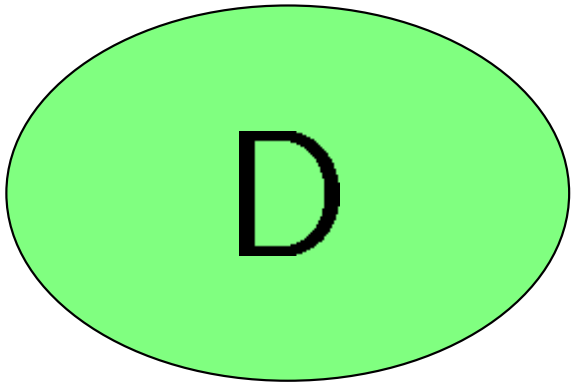
Tomasz Gogacz, Jerzy Marcinkowski

Uniwersytet Wrocławski

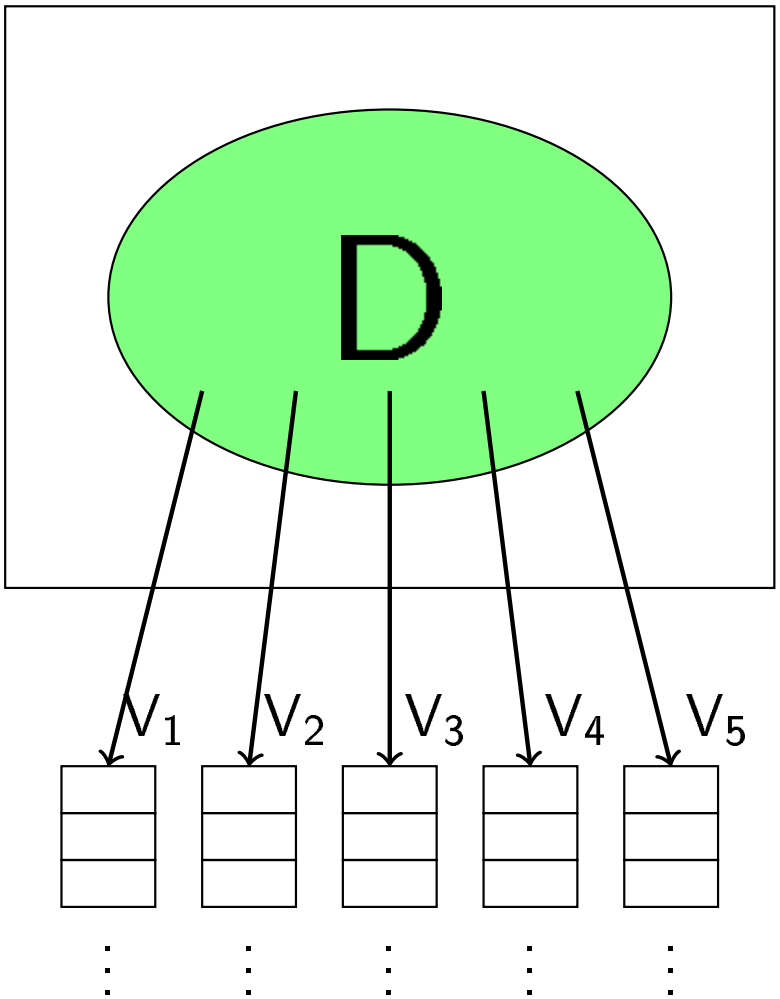
gogo@cs.uni.wroc.pl

30.1.2015

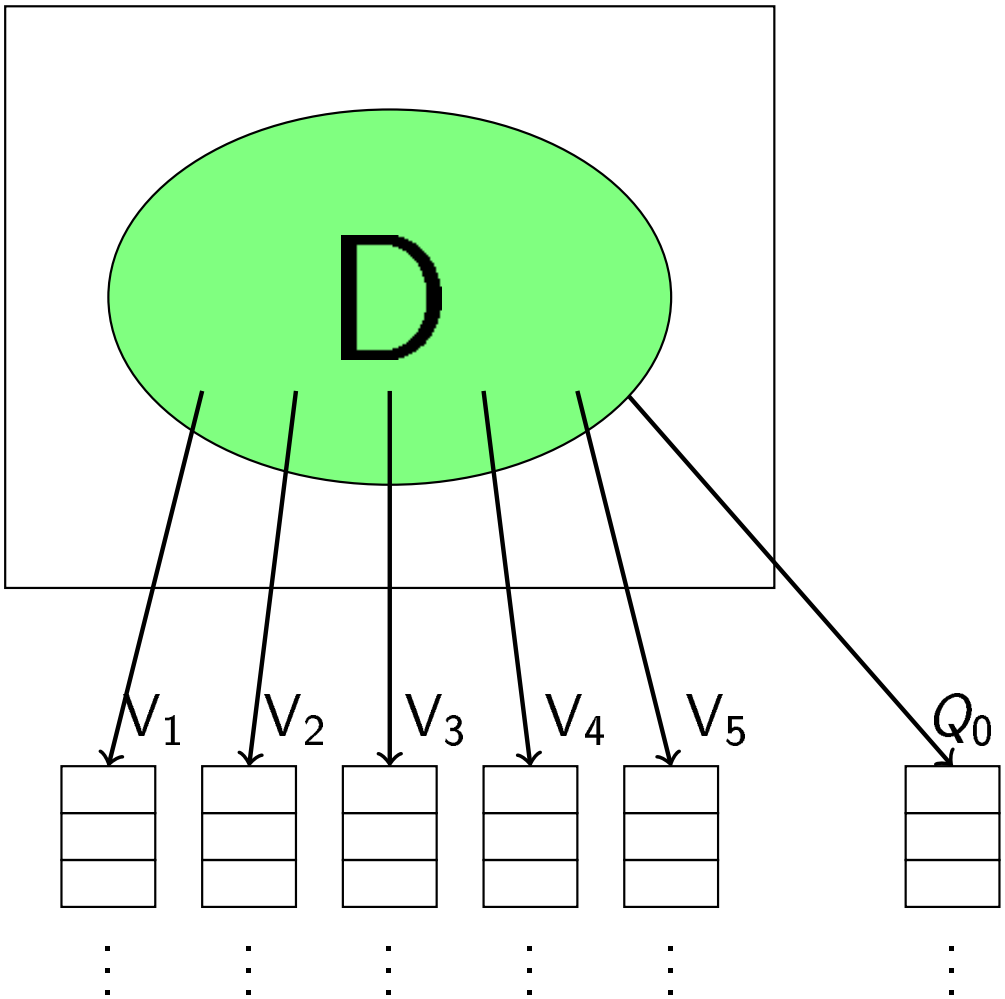
Framework



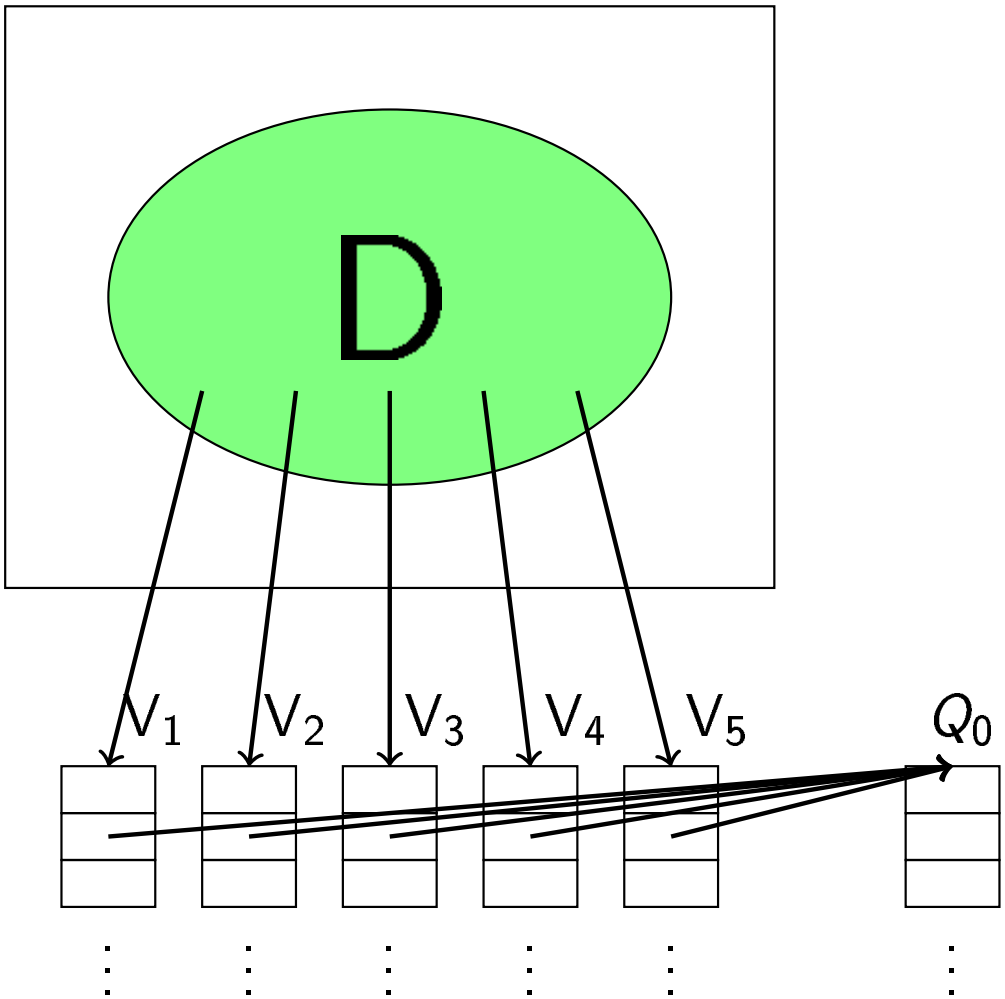
Framework



Framework



Framework



Determinacy

Definition of determinacy (informal)

We say that a set of views $\mathcal{V}$ determines a query Q if $\mathcal{V}$ always provides enough information to give answer to Q.

Determinacy

Definition of determinacy (informal)

We say that a set of views $\mathcal{V}$ determines a query Q if $\mathcal{V}$ always provides enough information to give answer to Q.

Definition of determinacy (formal)

We say that a set of views $\mathcal{V}$ determines a query Q if for each pair of databases D_1, D_2:

$$\big(\forall i \; V_i(D_1) = V_i(D_2)\big) \Rightarrow Q(D_1) = Q(D_2)$$

Determinacy

Definition of determinacy (informal)

We say that a set of views $\mathcal{V}$ determines a query Q if $\mathcal{V}$ always provides enough information to give answer to Q.

Definition of determinacy (formal)

We say that a set of views $\mathcal{V}$ determines a query Q if for each pair of databases D_1, D_2:

$$\left(\forall i \; V_i(D_1) = V_i(D_2)\right) \Rightarrow Q(D_1) = Q(D_2)$$

Suppose we have all tuples $T_i = V_i(D)$. Find (by brute force) D' such $\forall i \;\; V_i(D') = T_i$. Then compute $Q(D')$.

Determinacy

Definition of determinacy (informal)

We say that a set of views $\mathcal{V}$ determines a query Q if $\mathcal{V}$ always provides enough information to give answer to Q.

Definition of determinacy (formal)

We say that a set of views $\mathcal{V}$ determines a query Q if for each pair of databases D_1, D_2:

$$\left(\forall i \; V_i(D_1) = V_i(D_2)\right) \Rightarrow Q(D_1) = Q(D_2)$$

Suppose we have all tuples $T_i = V_i(D)$. Find (by brute force) D' such $\forall i \;\; V_i(D') = T_i$. Then compute $Q(D')$.

Rewriting

Q' is a rewriting of Q under $\mathcal{V}$ iff for all databases D the following equality holds $Q(D) = Q'(\mathcal{V}(D))$.

Our Theorem

Problem

Input: Set of views $\mathcal{V}$ and query Q.
Question: Does $\mathcal{V}$ determine Q?

Theorem

The problem above is undecidable when $\mathcal{V}$ is a set of conjunctive queries and Q is a conjunctive query.

Query Language

Conjunctive Queries

$$Q(\bar{x}) = \exists \bar{y}\ \Phi(\bar{x},\ \bar{y})$$

Examples

Path of lenght 3

$P_3(x, y) = \exists z_1, z_2\ E(x, z_1) \wedge E(z_1, z_2) \wedge E(z_2, y)$

In SQL

```
SELECT edge1.x, edge3.y
FROM edges edge1, edges edge2, edges edge3
WHERE edge1.y=edge2.x AND edge2.y=edge3.x;
```

Previous work

- 1985 P. Larson H. Yang, Computing queries from derived relations
- 1987 H. Yang P. Larson, Query transformation for psj-queries
- 1995 A. Levy A. Mendelzon Y. Sagiv, Answering queries using views
- 2005 L. Segoufin V. Vianu, Views and queries: Determinacy and rewriting
- 2007 F. Afrati, Determinacy and query rewriting for conjunctive queries and views
- 2011 D. Pasaila, Conjunctive queries determinacy and rewriting
- 2012 W. Fan F. Geerts L. Zheng, View determinacy for preserving selected information in data transformations.

Examples of (non-)determinacy

Path queries

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$

Does $\{P_3, P_4\}$ determine P_7?

Examples of (non-)determinacy

Path queries

$P_3(x, y) = \exists z_1, z_2 \quad E(x, z_1) \wedge E(z_1, z_2) \wedge E(z_2, y)$
$P_4(x, y) = \exists z_1, z_2, z_3 \; E(x, z_1) \wedge E(z_1, z_2) \wedge E(z_2, z_3) \wedge E(z_3, y)$

Does $\{P_3, P_4\}$ determine P_7? Yes

$$P_7(x, y) = \exists z \; P_3(x, z) \wedge P_4(z, y)$$

Examples of (non-)determinacy

Path queries

$P_3(x, y) = \exists z_1, z_2 \quad E(x, z_1) \wedge E(z_1, z_2) \wedge E(z_2, y)$
$P_4(x, y) = \exists z_1, z_2, z_3 \; E(x, z_1) \wedge E(z_1, z_2) \wedge E(z_2, z_3) \wedge E(z_3, y)$

Does $\{P_3, P_4\}$ determine P_7? Yes
Does $\{P_3, P_4\}$ determine P_2?

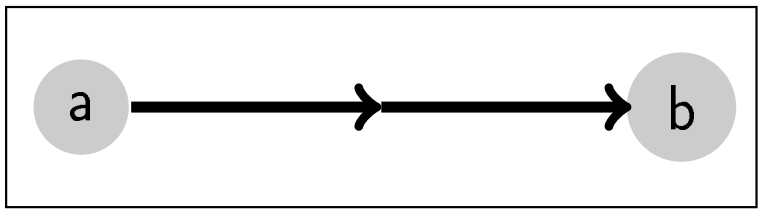
Examples of (non-)determinacy

Path queries

$P_3(x, y) = \exists z_1, z_2 \quad E(x, z_1) \wedge E(z_1, z_2) \wedge E(z_2, y)$
$P_4(x, y) = \exists z_1, z_2, z_3 \; E(x, z_1) \wedge E(z_1, z_2) \wedge E(z_2, z_3) \wedge E(z_3, y)$

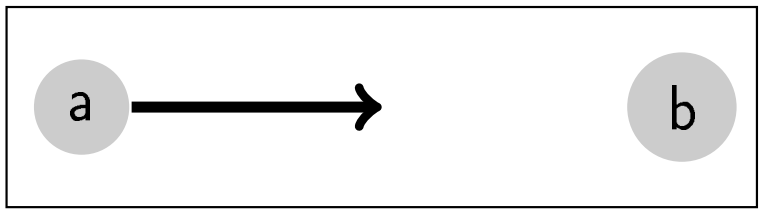
Does $\{P_3, P_4\}$ determine P_7? Yes
Does $\{P_3, P_4\}$ determine P_2? No

D_1



$P_3(D_1) = P_3(D_1) = \emptyset$
$P_2(D_1) = \langle a, b \rangle$

D_2



$P_3(D_2) = P_3(D_2) = \emptyset$
$P_2(D_2) = \emptyset$

Examples of (non-)determinacy

Path queries

$P_3(x, y) = \exists z_1, z_2 \quad E(x, z_1) \wedge E(z_1, z_2) \wedge E(z_2, y)$
$P_4(x, y) = \exists z_1, z_2, z_3 \; E(x, z_1) \wedge E(z_1, z_2) \wedge E(z_2, z_3) \wedge E(z_3, y)$

Does $\{P_3, P_4\}$ determine P_7? Yes
Does $\{P_3, P_4\}$ determine P_2? No
Does $\{P_3, P_4\}$ determine P_5?

How to get a counterexample

$\{P_3, P_4\}$ does not determine P_2

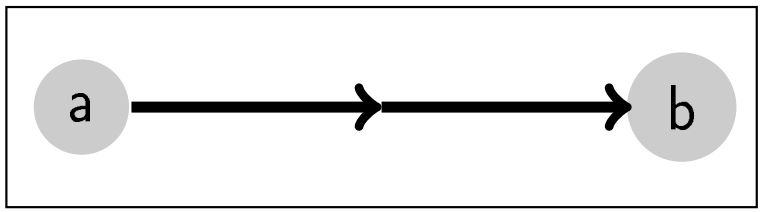
$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$

How to get a counterexample

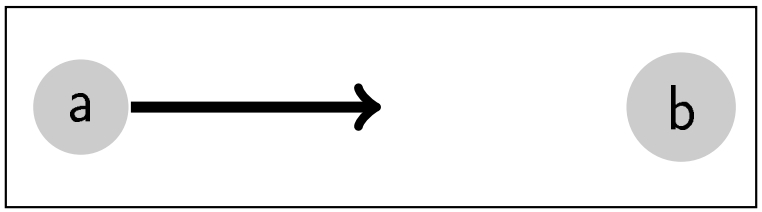
$\{P_3, P_4\}$ does not determine P_2

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$

D_1



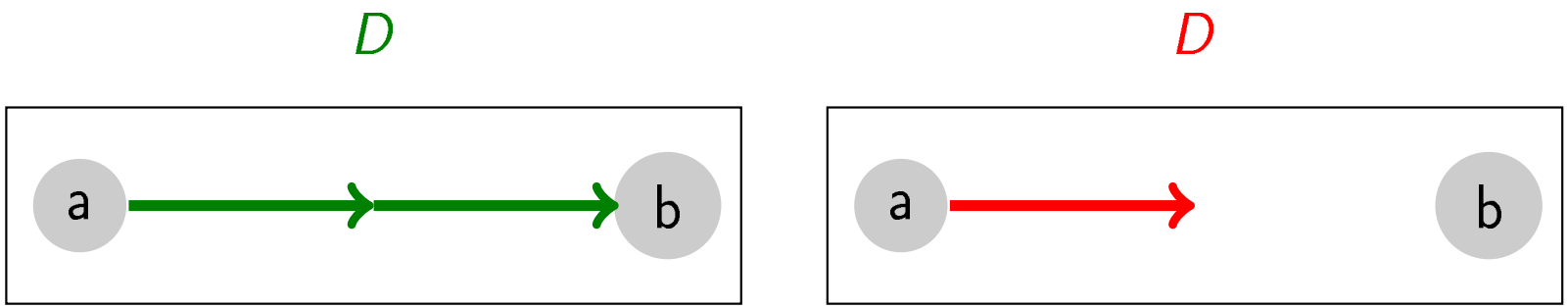
D_2



How to get a counterexample

$\{P_3, P_4\}$ does not determine P_2

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$

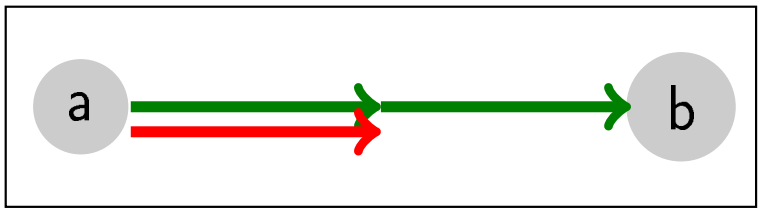


How to get a counterexample

$\{P_3, P_4\}$ does not determine P_2

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$

D

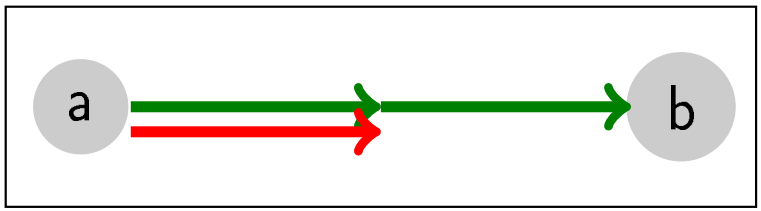


How to get a counterexample

$\{P_3, P_4\}$ does not determine P_2

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$

D



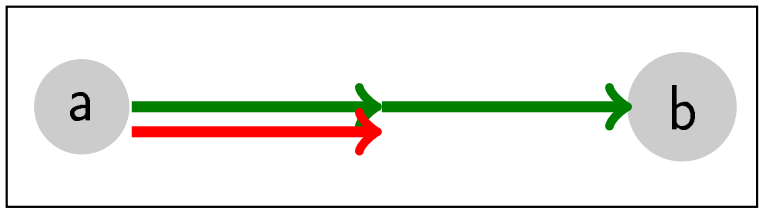
$$P_3(D_1) = P_3(D_2) \qquad D \models \forall x,y \;\; P_3(x,y) \rightarrow P_3(x,y)$$
$$D \models \forall x,y \;\; P_3(x,y) \rightarrow P_3(x,y)$$

How to get a counterexample

$\{P_3, P_4\}$ does not determine P_2

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$

D



$$P_3(D_1) = P_3(D_2) \qquad D \models \forall x,y \;\; P_3(x,y) \rightarrow P_3(x,y)$$
$$D \models \forall x,y \;\; P_3(x,y) \rightarrow P_3(x,y)$$

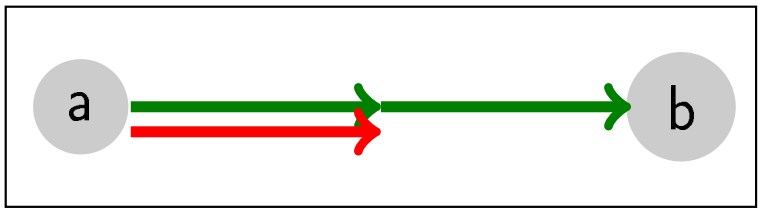
$$\forall x, y, z_1, z_2 \; E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y) \;\Rightarrow\; \exists w_1, w_2 \; E(x,w_1) \wedge E(w_1,w_2) \wedge E(w_2,y)$$

How to get a counterexample

$\{P_3, P_4\}$ does not determine P_2

$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$
$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$

D



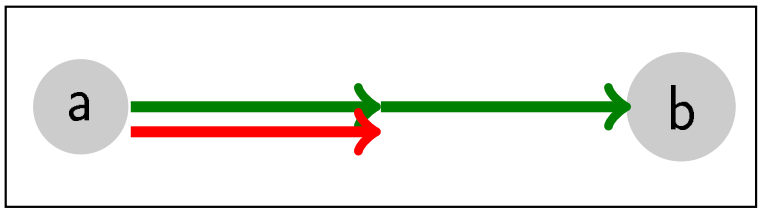
$P_3(D_1) = P_3(D_2)$ $\quad D \models \forall x,y \;\; P_3(x,y) \rightarrow P_3(x,y)$
$\quad D \models \forall x,y \;\; P_3(x,y) \rightarrow P_3(x,y)$
$P_4(D_1) = P_4(D_2)$ $\quad D \models \forall x,y \;\; P_4(x,y) \rightarrow P_4(x,y)$
$\quad D \models \forall x,y \;\; P_4(x,y) \rightarrow P_4(x,y)$

How to get a counterexample

$\{P_3, P_4\}$ does not determine P_2

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$

D



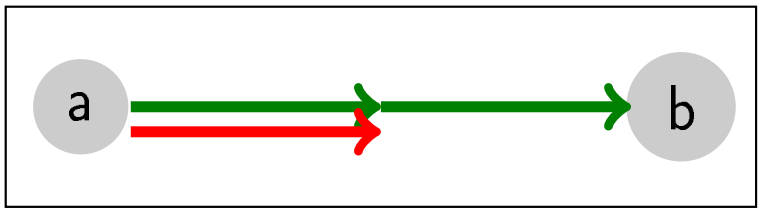
$$P_3(D_1) = P_3(D_2) \quad D \models \forall x,y \;\; P_3(x,y) \rightarrow P_3(x,y)$$
$$D \models \forall x,y \;\; P_3(x,y) \rightarrow P_3(x,y)$$
$$P_4(D_1) = P_4(D_2) \quad D \models \forall x,y \;\; P_4(x,y) \rightarrow P_4(x,y)$$
$$D \models \forall x,y \;\; P_4(x,y) \rightarrow P_4(x,y)$$
$$D \models P_2(a,b) \wedge \neg P_2(a,b)$$

How to get a counterexample

$\{P_3, P_4\}$ does not determine P_2

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$

D



Formulas in $\mathscr{V}$

$$D \models \forall x,y \quad P_3(x,y) \rightarrow P_3(x,y)$$
$$D \models \forall x,y \quad P_3(x,y) \rightarrow P_3(x,y)$$
$$D \models \forall x,y \quad P_4(x,y) \rightarrow P_4(x,y)$$
$$D \models \forall x,y \quad P_4(x,y) \rightarrow P_4(x,y)$$

$$D \models P_2(a,b) \wedge \neg P_2(a,b)$$

Problem restated

Problem

Input: Set of views $\mathcal{V}$ and query Q.

Question: Does $\mathcal{V}, Q \models Q$?

Tuple Generating Dependencies

$$\forall\, \bar{x}\bar{y}\ \Phi(\bar{x}, \bar{y}) \ \rightarrow\ \exists \bar{z}\ \Psi(\bar{x}, \bar{z})$$

Tuple Generating Dependencies

$$\forall\, \bar{x}\bar{y}\ \Phi(\bar{x},\bar{y}) \ \rightarrow\ \exists \bar{z}\ \Psi(\bar{x},\bar{z})$$

Existence of a universal structure

$Q, \mathcal{V} \models Q \quad \Leftrightarrow \quad Chase(Q, \mathcal{V}) \models Q$

Tuple Generating Dependencies

$$\forall\, \bar{x}\bar{y}\ \Phi(\bar{x},\bar{y}) \ \rightarrow\ \exists \bar{z}\ \Psi(\bar{x},\bar{z})$$

Existence of a universal structure

$Q, \mathcal{W} \models Q \quad \Leftrightarrow \quad Chase(Q, \mathcal{W}) \models Q$

Chase construction

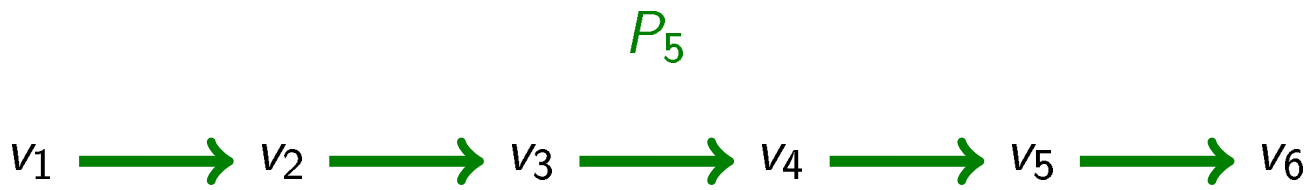
It is a fixpoint. Whenever you see a copy of $\Phi(\bar{x},\bar{y})$ create a new copy of $\Psi(\bar{x},\bar{z})$.

Why $\{P_3, P_4\}$ determines P_5

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$

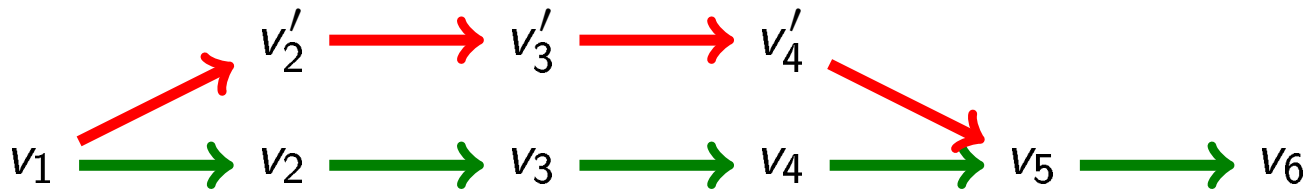
Why $\{P_3, P_4\}$ determines P_5

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$



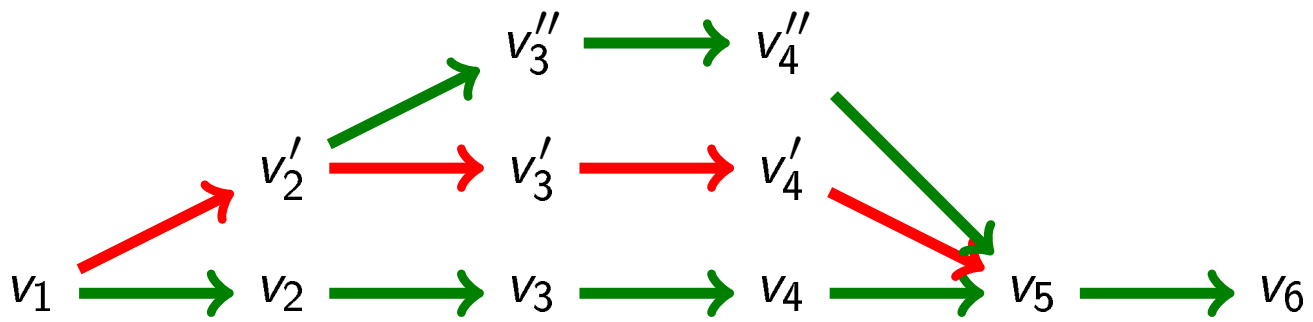
Why $\{P_3, P_4\}$ determines P_5

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$



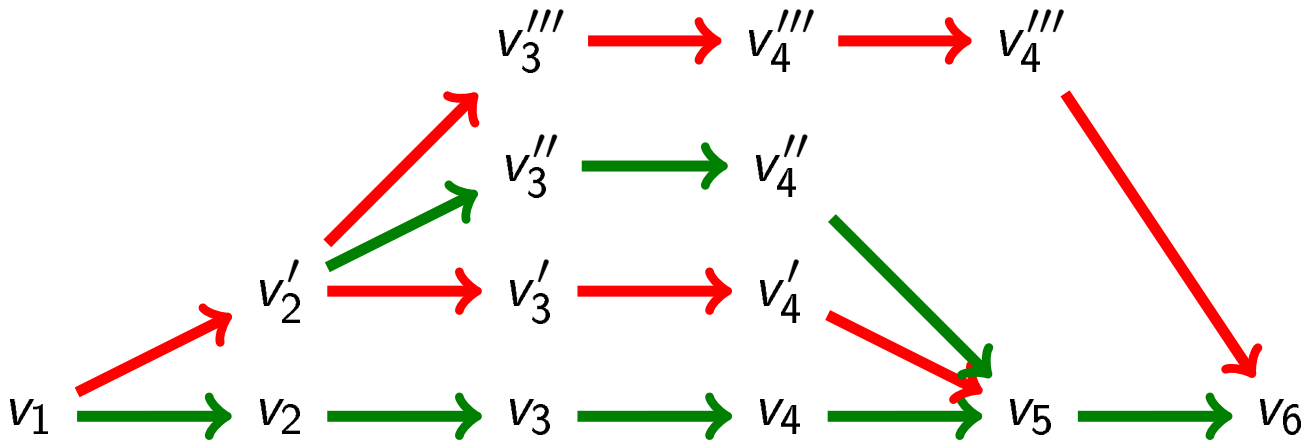
Why $\{P_3, P_4\}$ determines P_5

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$



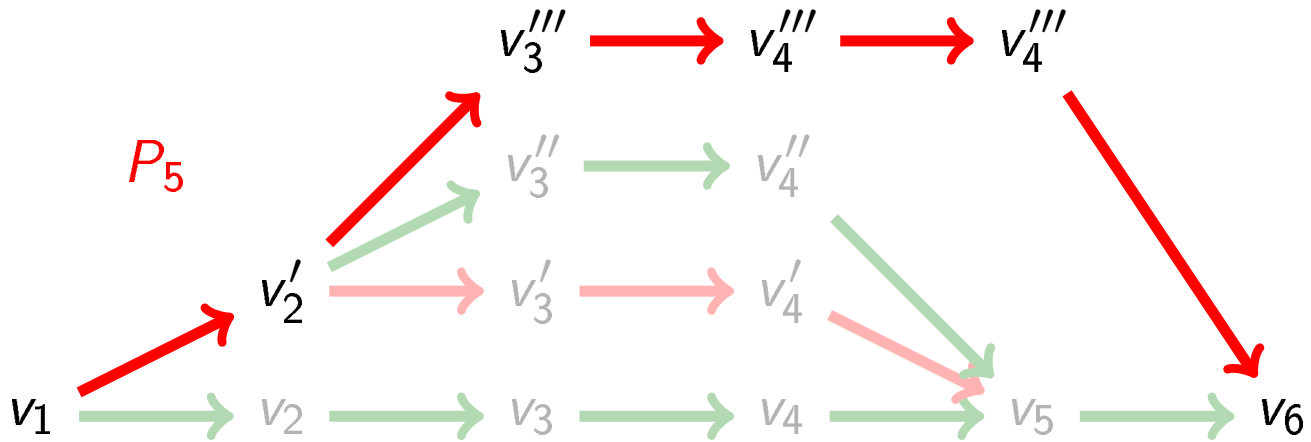
Why $\{P_3, P_4\}$ determines P_5

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$



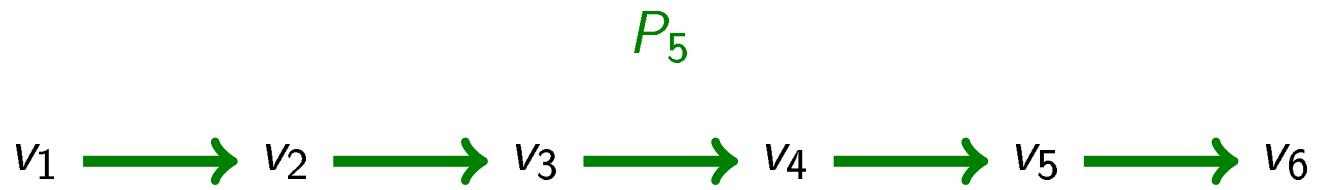
Why $\{P_3, P_4\}$ determines P_5

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$



Why $\{P_3, P_4\}$ determines P_5

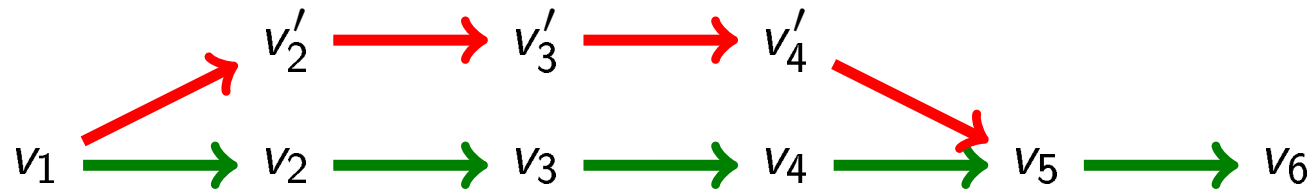
$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$



$P_5(x,y) =$

Why $\{P_3, P_4\}$ determines P_5

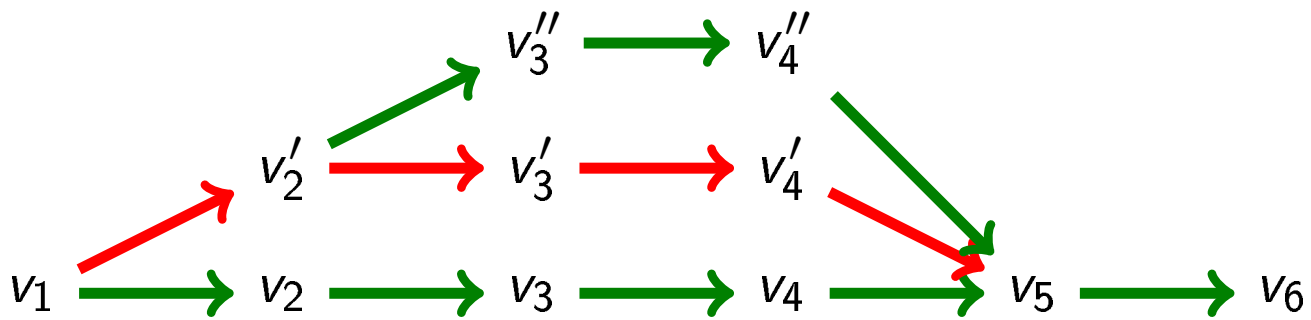
$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$



$$P_5(x,y) = \exists v_5 \; P_4(x,v_5)$$

Why $\{P_3, P_4\}$ determines P_5

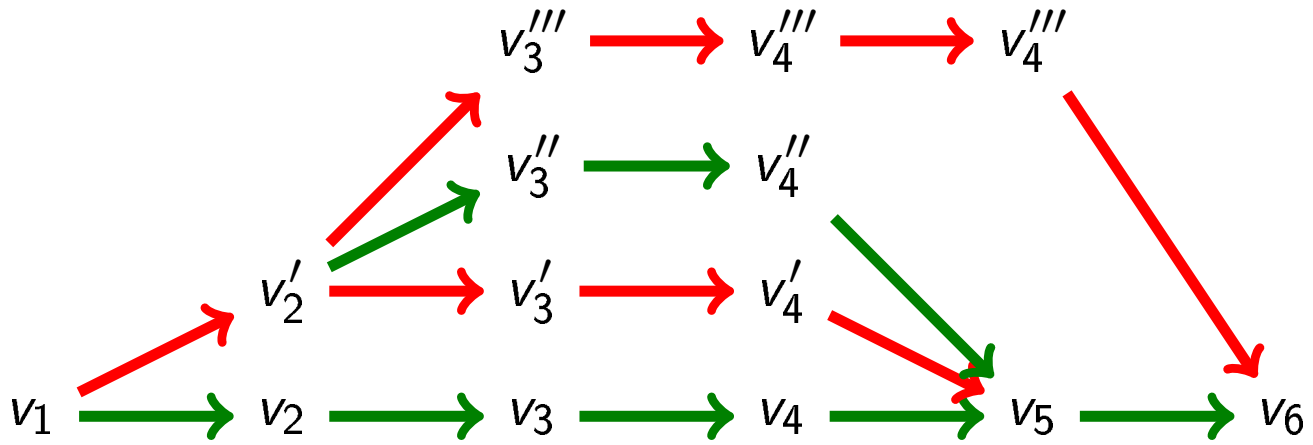
$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \; E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$



$$P_5(x,y) = \exists v_5 \; P_4(x,v_5) \wedge (\forall v_2 \; P_3(v_2,v_5)$$

Why $\{P_3, P_4\}$ determines P_5

$$P_3(x,y) = \exists z_1, z_2 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,y)$$
$$P_4(x,y) = \exists z_1, z_2, z_3 \quad E(x,z_1) \wedge E(z_1,z_2) \wedge E(z_2,z_3) \wedge E(z_3,y)$$



$$P_5(x,y) = \exists v_5 \; P_4(x,v_5) \wedge (\forall v_2 \; P_3(v_2,v_5) \Rightarrow P_4(v_2,y))$$

Main obstacle

Lemma

If Chase$(D, \mathscr{V}) \models Q$ then $dalt(D) \models dalt(Q)$