View-based query processing

Reasoning about views

Diego Calvanese, Giuseppe De Giacomo,

Maurizio Lenzerini, Riccardo Rosati, Georg Gottlob

Corso di dottorato – Dottorato in Ingegneria Informatica,

Università di Roma "La Sapienza", 3 novembre 2005

View based query processing

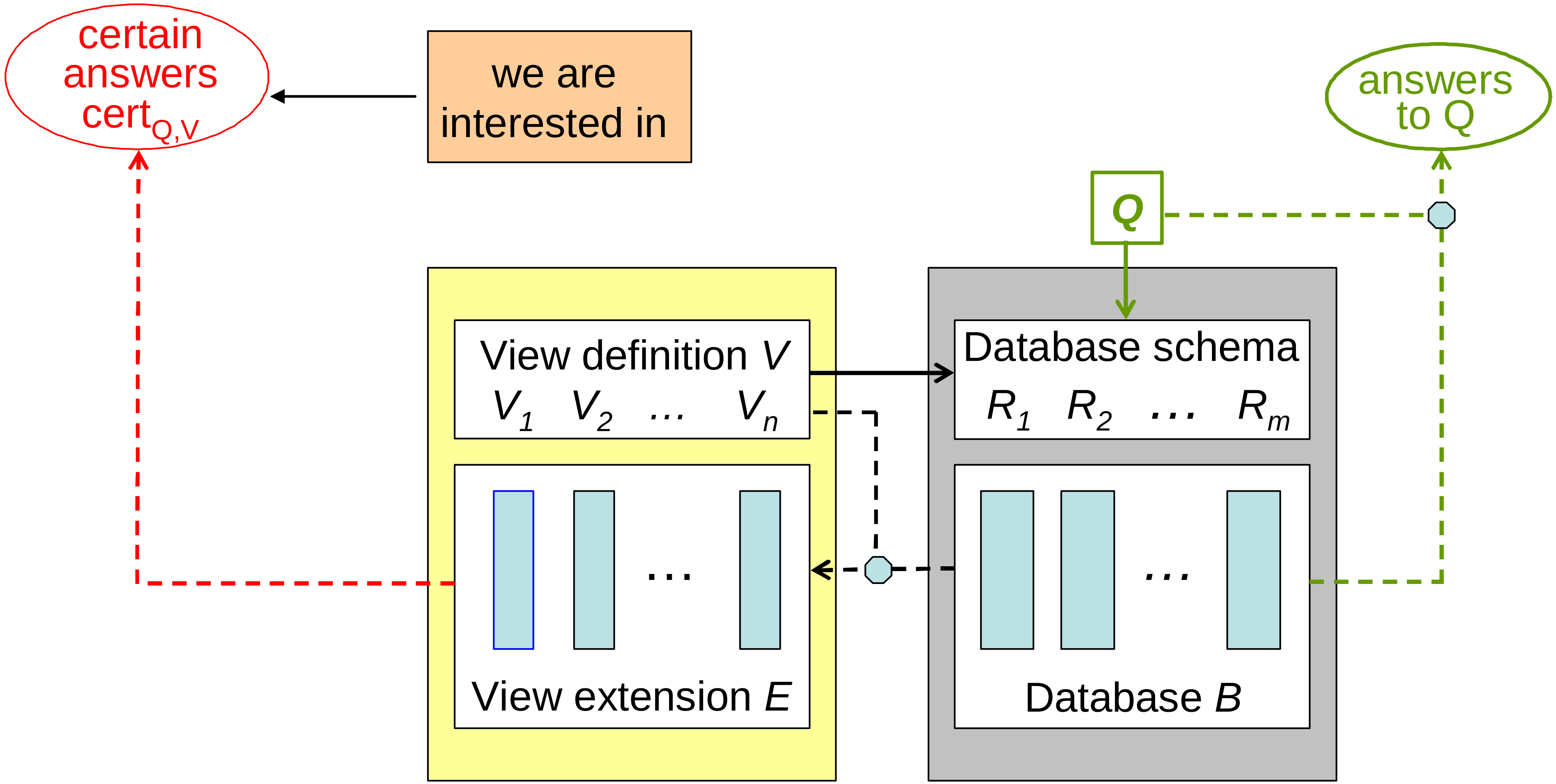
Computing the answer to a query by relying solely on a set of views

Relevant problem in data integration, data warehousing, query optimization, authorization, etc.

Two different approaches:

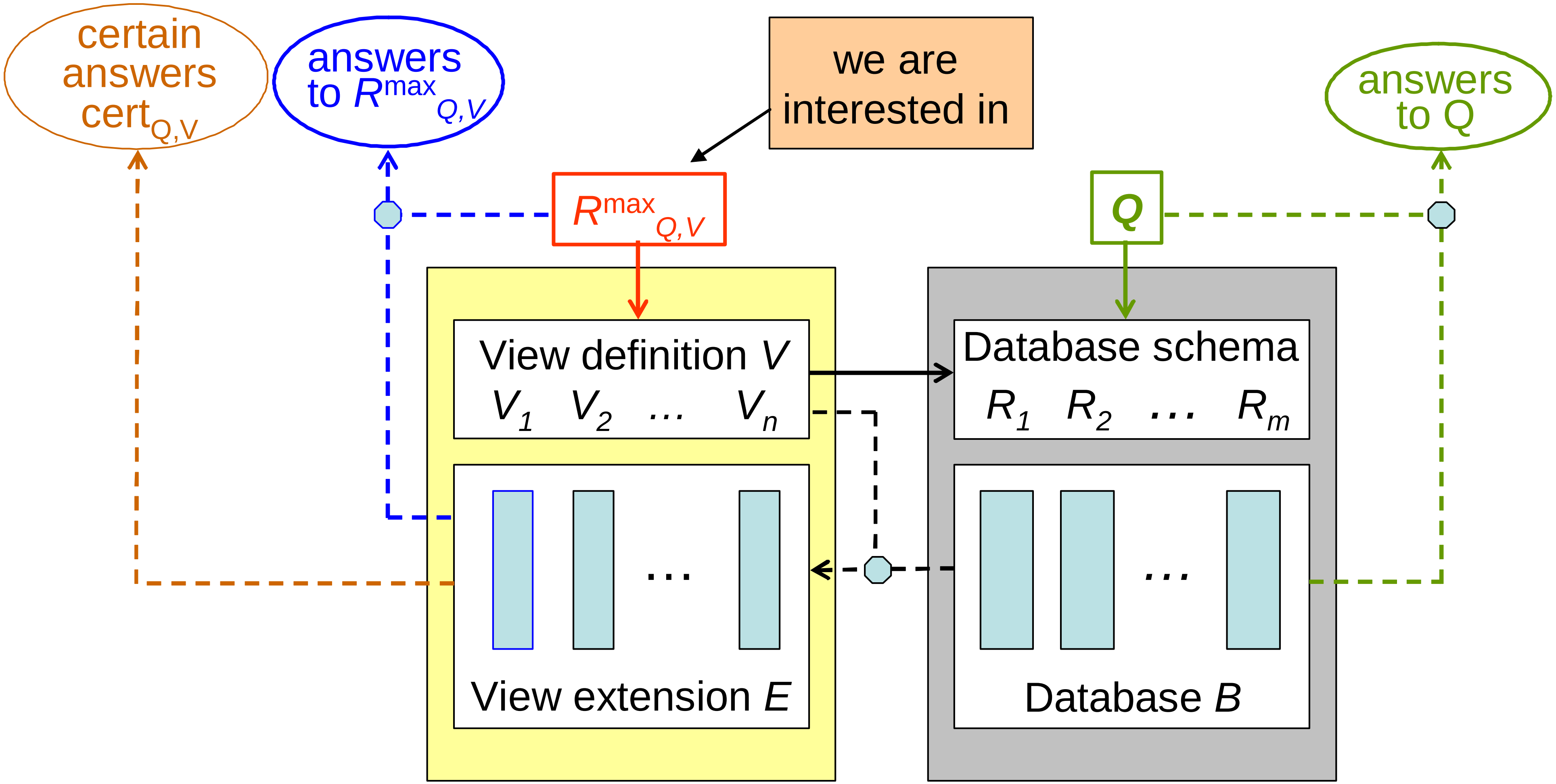
- view based query answering
- view based query rewriting

View based query answering



Open world assumption (sound views): $\mathcal{E} \subseteq \mathcal{V}(\mathcal{B})$

View based query rewriting



Open world assumption (sound views): $\mathcal{E} \subseteq \mathcal{V}(\mathcal{B})$

$R^{max}_{Q,\mathcal{V}}$ expressed in the “same” language as Q (but on $\mathcal{V}$-symbols)

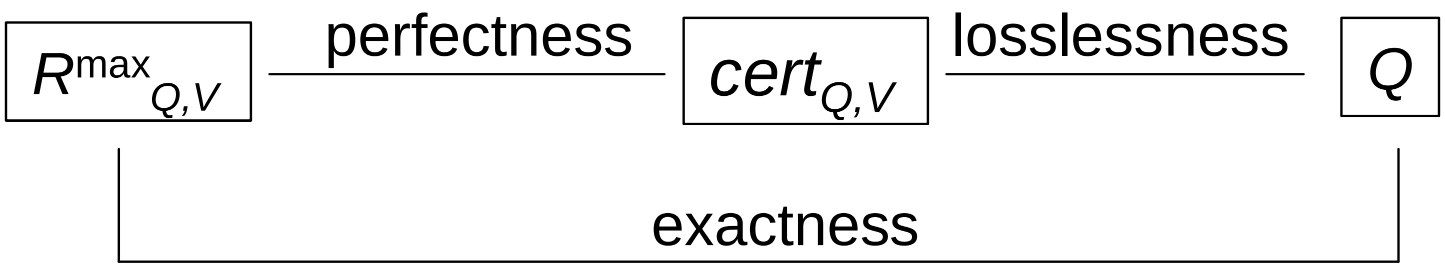
Answering vs rewriting

- Answering and rewriting coincide in some interesting cases (notably, in the case of conjunctive queries and views – see later)
- However, they do not coincide in general, and therefore, it makes sense to compare the query, the rewriting and the certain answers

The main focus of this lecture

Principles and tools for comparing:

- query Q
- maximal rewriting $R^{max}_{Q,\mathcal{V}}$ of Q wrt views $\mathcal{V}$
  (a maximal rewriting of Q wrt $\mathcal{V}$ is a maximal query R over $\mathcal{V}$ such that $\forall \mathcal{B} \forall \mathcal{E} \subseteq \mathcal{V}(\mathcal{B})$: we have $R(\mathcal{E}) \subseteq Q(\mathcal{B})$)
- function (i.e., query) $cert_{Q,\mathcal{V}}$ that computes the certain answers to Q wrt views $\mathcal{V}$, given $\mathcal{V}$-extension $\mathcal{E}$
  (i.e., $\vec{\mathbf{t}} \in cert_{Q,\mathcal{V}}(\mathcal{E})$ iff $\forall \mathcal{B} : \mathcal{E} \subseteq \mathcal{V}(\mathcal{B})$ we have $\vec{\mathbf{t}} \in Q(\mathcal{B})$)



Outline

1. Framework
2. Rewriting vs answering
3. Exactness
4. Perfectness
5. Losslessness
6. Conclusions

The lecture is based on the paper:

[CDLV05] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Moshe Y. Vardi. "View-Based Query Processing: On the Relationship Between Rewriting, Answering and Losslessness". *Proc. of the International Conference on Database Theory, ICDT 2005*

Framework

Two settings:

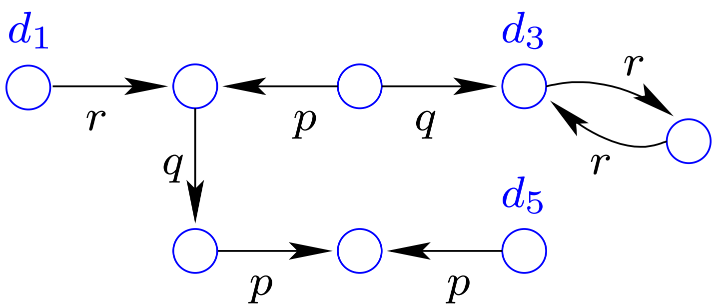
- Relational dbs
  - Conjunctive queries and views
- Semistructured data: edge labeled graph, with set Σ of basic binary relations (edge labels) on nodes
  - Queries and views: variants of regular path queries (RPQs)
    - $*$ an RPQ is a regular expression Q over the edge labels
    - $*$ it returns the set of pairs of nodes connected by a path in $L(Q)$

Two-way regular path queries (2RPQs)

Expressed as regular expression over $\Sigma^{\pm} = \Sigma \cup \{p^- \mid p \in \Sigma\}$ (p^- denotes the inverse of the binary relation p), e.g.,

$$Q = r \cdot (p^- + q) \cdot p \cdot p^- \cdot q^*$$

Answer over DB: set of pairs of nodes connected by a semipath in DB conforming to the regular expression

DB: 

Q returns

d_1	d_5
d_1	d_3
$\vdots$	$\vdots$

via $rqpp^-$
via rp^-pp^-q

Rewriting vs answering for conjunctive queries

$$
\begin{array}{lcl}
\mathsf{v}_1(T) & = & \{\ (T) \mid \mathsf{movie}(T,Y,D) \wedge \mathsf{european}(D)\ \} \\
\mathsf{v}_2(T,Z) & = & \{\ (T,Z) \mid \mathsf{movie}(T,Y,D) \wedge \mathsf{review}(T,Z)\ \}
\end{array}
$$

The certain answers to Q are computed by evaluating the goal Q wrt this nonrecursive logic program [Abiteboul&Duschka 1998]:

$$
\begin{array}{rcl}
\mathsf{movie}(T, f_1(T), f_2(T)) & \leftarrow & \mathsf{v}_1(T) \\
\mathsf{european}(f_2(T)) & \leftarrow & \mathsf{v}_1(T) \\
\mathsf{movie}(T, f_4(T,Z), f_5(T,Z)) & \leftarrow & \mathsf{v}_2(T,Z) \\
\mathsf{review}(T,Z)) & \leftarrow & \mathsf{v}_2(T,Z)
\end{array}
$$

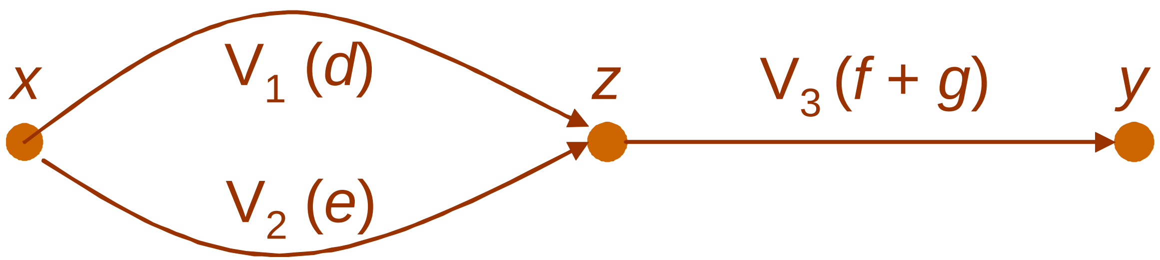
The goal and the logic program can be equivalently transformed into a finite union of conjunctive queries over the view symbols, which is the maximal rewriting of Q wrt $\mathcal{V}$

Rewriting vs answering for 2RPQs

Views $\mathcal{V}$:	$V_1 = d \quad V_2 = e \quad V_3 = f + g$
Query Q:	$df + eg$

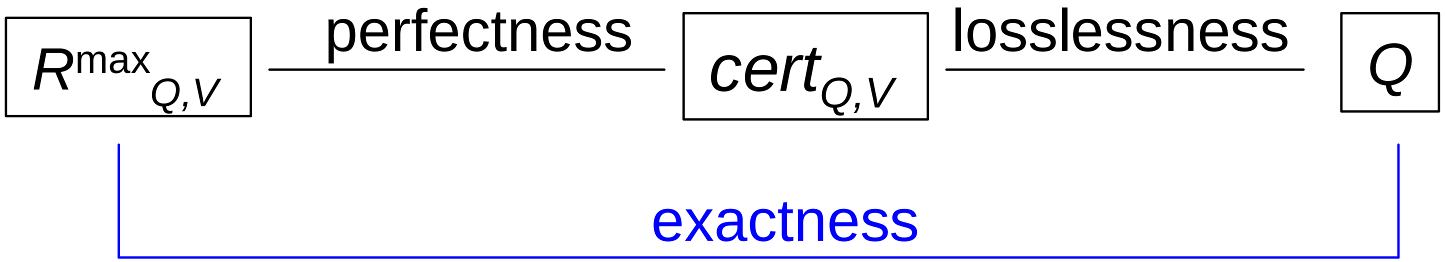
$R^{max}_{Q,\mathcal{V}} = \emptyset$

$$cert_{Q,\mathcal{V}} = \{ \ (x, y) \mid \exists z.\ x\, V_1\, z \wedge x\, V_2\, z \ \wedge z\, V_3\, y \ \}$$



Furthermore, computing the certain answers is coNP-complete in data complexity, while evaluating the maximal rewriting can be done in polynomial time

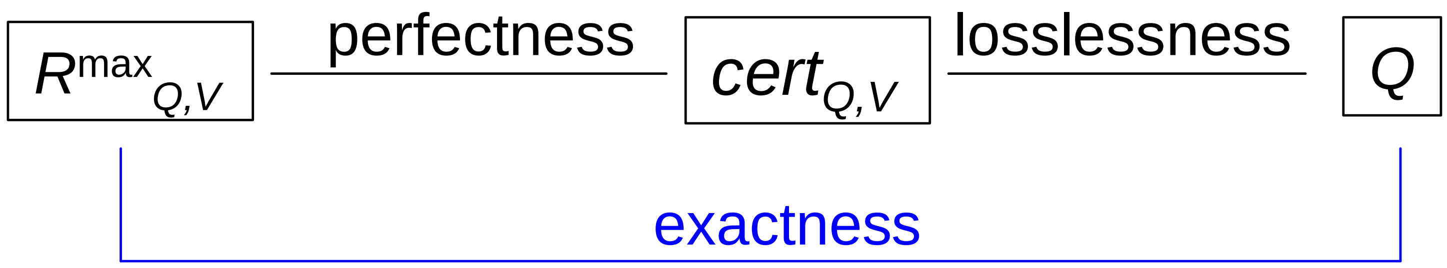
Exactness: comparing $R^{max}_{Q,\mathcal{V}}$ and Q



The maximal rewriting $R^{max}_{Q,\mathcal{V}}$ of Q wrt views $\mathcal{V}$ is exact if for every database $\mathcal{B}$ we have that $Q(\mathcal{B}) = R^{max}_{Q,\mathcal{V}}(\mathcal{V}(\mathcal{B}))$

Exactness means losslessness of rewriting wrt the query (note that exactness = perfectness + losslessness)

Exactness in the case of conjunctive queries



- $R^{max}_{Q,\mathcal{V}}$ is a union of conjunctive queries over the $\mathcal{V}$-symbols
- To check whether such union is equivalent to Q modulo $\mathcal{V}$, it suffices to check whether there is a disjunct in the unfolding of $R^{max}_{Q,\mathcal{V}}$ that is equivalent to Q
- Checking whether there exists an exact rewriting of a conjunctive query is NP-complete [Halevy&al 1995]

Exactness in the case of 2RPQs

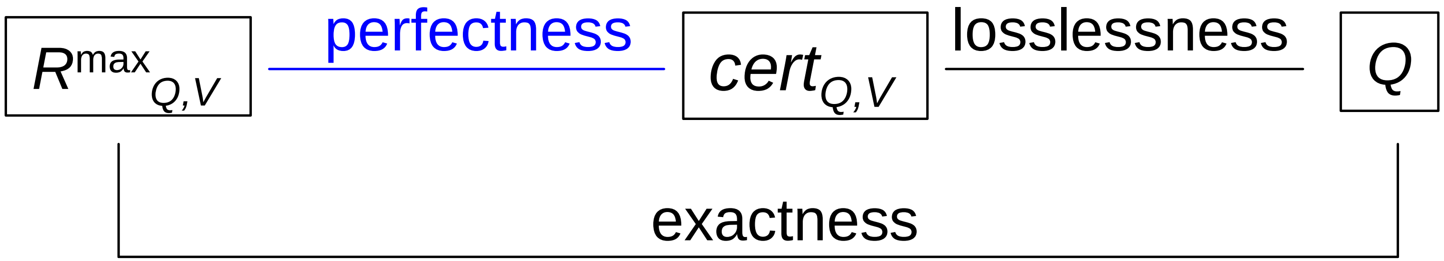
$\boxed{R^{max}_{Q,V}}$ $\xrightarrow{\text{perfectness}}$ $\boxed{cert_{Q,V}}$ $\xrightarrow{\text{losslessness}}$ $\boxed{Q}$

exactness

From [Calvanese&al 2000]:

- $R^{max}_{Q,\mathcal{V}}$ can be constructed in 2EXPTIME, via an automata-theoretic approach
- To check exactness, we check whether Q is contained in the unfolding of $R^{max}_{Q,\mathcal{V}}$
- Checking whether there exists an exact rewriting of a 2RPQ is 2EXPSPACE-complete

Perfectness: comparing $R^{max}_{Q,\mathcal{V}}$ and $cert_{Q,\mathcal{V}}$

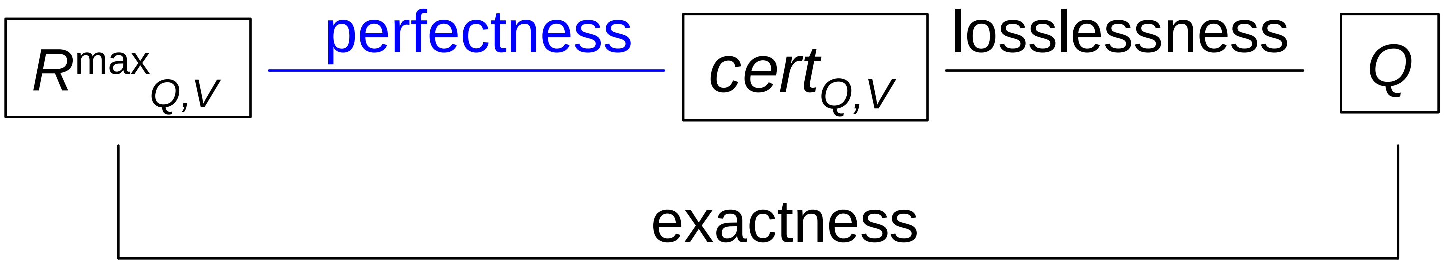


The maximal rewriting $R^{max}_{Q,\mathcal{V}}$ of Q wrt views $\mathcal{V}$ is perfect, if for every database $\mathcal{B}$ and every view extension $\mathcal{E}$ with $\mathcal{E} \subseteq \mathcal{V}(\mathcal{B})$ we have that $cert_{Q,\mathcal{V}}(\mathcal{E}) = R^{max}_{Q,\mathcal{V}}(\mathcal{E})$

Perfectness means that the maximal rewriting is powerful enough to compute the certain answers

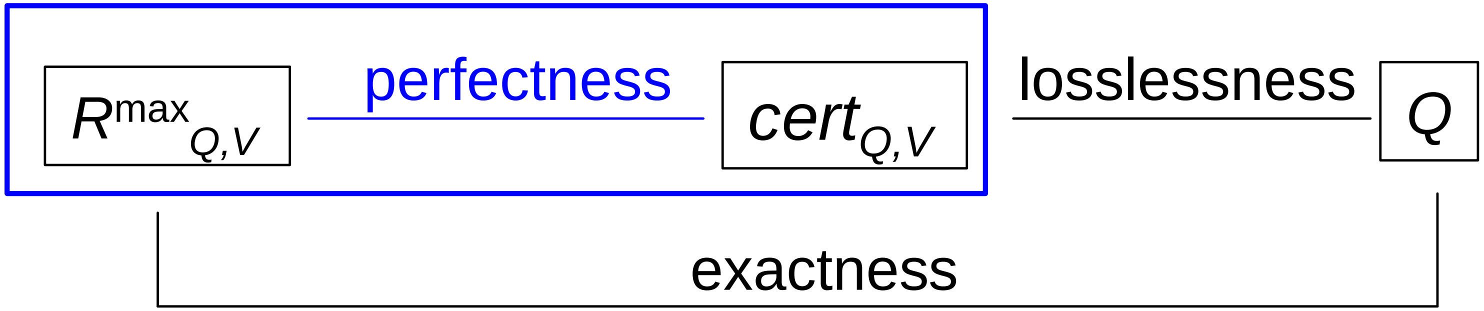
If $R^{max}_{Q,\mathcal{V}}$ is perfect, then we can compute $cert_{Q,\mathcal{V}}$ by evaluating $R^{max}_{Q,\mathcal{V}}$ over the view extension

Perfectness in the case of conjunctive queries



What can we say about perfectness in the case of conjunctive queries?

Perfectness in the case of conjunctive queries



- $R^{max}_{Q,\mathcal{V}}$ is always equivalent to $cert_{Q,\mathcal{V}}$
- $R^{max}_{Q,\mathcal{V}}$ is always perfect

Perfectness in the case of 2RPQs

Perfectness means

$$\forall \mathcal{B}\ \forall \mathcal{E} \subseteq \mathcal{V}(\mathcal{B}) :\ cert_{Q,\mathcal{V}}(\mathcal{E}) \subseteq R^{max}_{Q,\mathcal{V}}(\mathcal{E})$$

that is a form of view-based query containment

$$Q \subseteq_{\mathcal{V}} R^{max}_{Q,\mathcal{V}}$$

View-based query containment is the problem of checking containment of two queries relative to set of views

Example of View-based Containment

Virtual schema: $Person(pname, worksfor, livesin)$
$Company(cname, budget)$
$European(nation, inhabitants)$

Queries:

$$\mathbf{Q_1}(p) \leftarrow Person(p, c, \_), Company(c, \_)$$
$$\mathbf{Q_2}(p) \leftarrow Person(p, \_, n), European(n, \_)$$

We have that $\mathbf{Q_1}$ is not contained in $\mathbf{Q_2}$ and $\mathbf{Q_2}$ is not contained in $\mathbf{Q_1}$.

Suppose data is only accessible through the view

$$\mathbf{V}(p, c) \leftarrow Person(p, c, n), Company(c, \_), European(n, \_)$$

Considering the data in the view only, $\mathbf{Q_1}$ and $\mathbf{Q_2}$ are indistinguishable.

View-based Containment – 4 Cases

We are given:

- an alphabet $\mathbf{\Sigma}$ of virtual relation symbols (base alphabet)
- an alphabet $\mathcal{V}$ of view symbols
- for each view $\mathbf{V}$ in $\mathcal{V}$, its definition, i.e., a query $\mathbf{V}^{\mathbf{\Sigma}}$ over $\mathbf{\Sigma}$
- two queries $\mathbf{Q_1}$ and $\mathbf{Q_2}$, each one either over $\mathbf{\Sigma}$ or over $\mathcal{V}$

We want to check whether $\mathbf{Q_1}$ is contained in $\mathbf{Q_2}$ relative to $\mathcal{V}$.

4 different cases, depending on the alphabet over which $\mathbf{Q_1}$ and $\mathbf{Q_2}$ are expressed:

$$1) \quad \mathbf{Q_1^{\Sigma}} \subseteq_{\mathcal{V}} \mathbf{Q_2^{\Sigma}} \qquad\qquad 2) \quad \mathbf{Q_1^{\mathcal{V}}} \subseteq_{\mathcal{V}} \mathbf{Q_2^{\Sigma}}$$

$$3) \quad \mathbf{Q_1^{\Sigma}} \subseteq_{\mathcal{V}} \mathbf{Q_2^{\mathcal{V}}} \qquad\qquad 4) \quad \mathbf{Q_1^{\mathcal{V}}} \subseteq_{\mathcal{V}} \mathbf{Q_2^{\mathcal{V}}}$$

Semantics of View-based Containment

$$1) \quad \mathbf{Q_1^\Sigma} \subseteq_{\mathcal{V}} \mathbf{Q_2^\Sigma} \qquad\qquad 2) \quad \mathbf{Q_1^{\mathcal{V}}} \subseteq_{\mathcal{V}} \mathbf{Q_2^\Sigma}$$

$$3) \quad \mathbf{Q_1^\Sigma} \subseteq_{\mathcal{V}} \mathbf{Q_2^{\mathcal{V}}} \qquad\qquad 4) \quad \mathbf{Q_1^{\mathcal{V}}} \subseteq_{\mathcal{V}} \mathbf{Q_2^{\mathcal{V}}}$$

if for every database $\mathcal{B}$, and for every $\mathcal{V}$-extension $\mathcal{E}$ with $\mathcal{E} \subseteq \mathcal{V}^{\mathbf{\Sigma}}(\mathcal{B})$, we have

$$1) \quad cert_{\mathbf{Q_1^\Sigma}}(\mathcal{E}) \subseteq cert_{\mathbf{Q_2^\Sigma}}(\mathcal{E}) \qquad\qquad 2) \quad \mathbf{Q_1^{\mathcal{V}}}(\mathcal{E}) \subseteq cert_{\mathbf{Q_2^\Sigma}}(\mathcal{E})$$

$$3) \quad cert_{\mathbf{Q_1^\Sigma}}(\mathcal{E}) \subseteq \mathbf{Q_2^{\mathcal{V}}}(\mathcal{E}) \qquad\qquad 4) \quad \mathbf{Q_1^{\mathcal{V}}}(\mathcal{E}) \subseteq \mathbf{Q_2^{\mathcal{V}}}(\mathcal{E})$$

Perfectness in the case of 2RPQs (cont.)

Perfectness

$$\forall \mathcal{B}\ \forall \mathcal{E} \subseteq \mathcal{V}(\mathcal{B}) : \ cert_{Q,\mathcal{V}}(\mathcal{E}) \subseteq R^{max}_{Q,\mathcal{V}}(\mathcal{E})$$

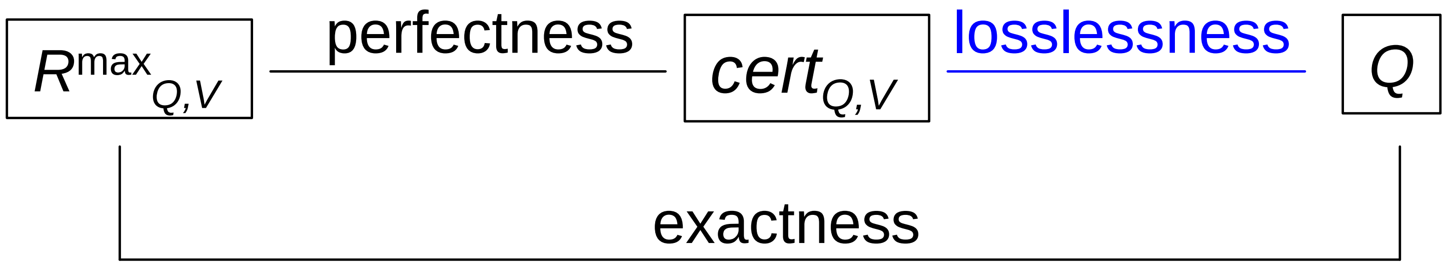
is therefore a form of view-based query containment

$$Q \subseteq_{\mathcal{V}} R^{max}_{Q,\mathcal{V}}$$

From [Calvanese&al 2003], this can be checked in NEXPTIME, resulting in N3EXPTIME for perfectness, since the size of $R^{max}_{Q,\mathcal{V}}$ is doubly exponential in Q

Actually, [CDLV05] shows that checking perfectness can be done in N2EXPTIME, via characterization of view-based query answering through CSP (lower bound open)

Losslessness: comparing $cert_{Q,\mathcal{V}}$ and Q



A set of views $\mathcal{V}$ is lossless wrt a query Q, if for every database $\mathcal{B}$ we have that $Q(\mathcal{B}) = cert_{Q,\mathcal{V}}(\mathcal{V}(\mathcal{B}))$

Losslessness means that the views are powerful enough to precisely answer the query

In the case where we have access to $\mathcal{B}$, losslessness allows us to compute $cert_{Q,\mathcal{V}}$ by evaluating Q over the database

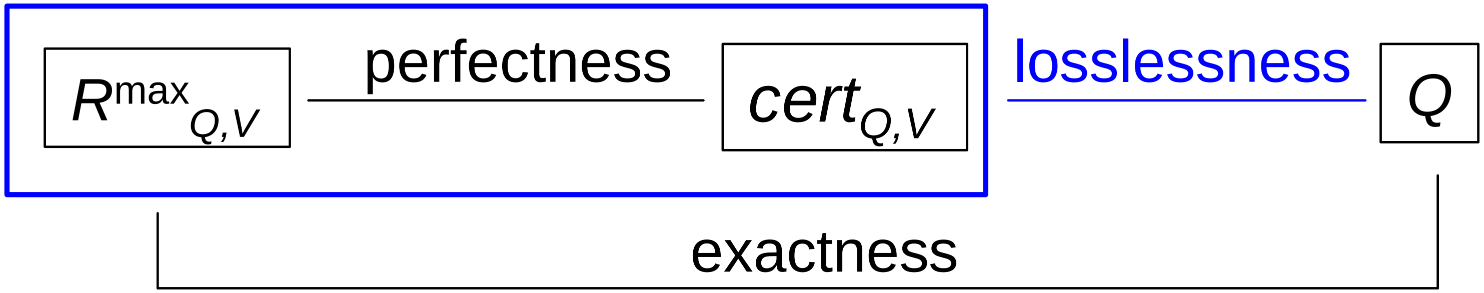
Losslessness in the case of conjunctive queries

$\boxed{R^{max}_{Q,V}}$ — perfectness — $\boxed{cert_{Q,V}}$ — losslessness — $\boxed{Q}$

$\underbrace{R^{max}_{Q,V} \quad \cdots \quad Q}_{\text{exactness}}$

What can we say about losslessness in the case of conjunctive queries?

Losslessness in the case of conjunctive queries



- $R^{max}_{Q,\mathcal{V}}$ is always equivalent to $cert_{Q,\mathcal{V}}$
- Losslessness and exactness coincide, i.e., checking losslessness is NP-complete

Losslessness in the case of 2RPQs: example

Views $\mathcal{V}$: $V_1 = 0 + 1$

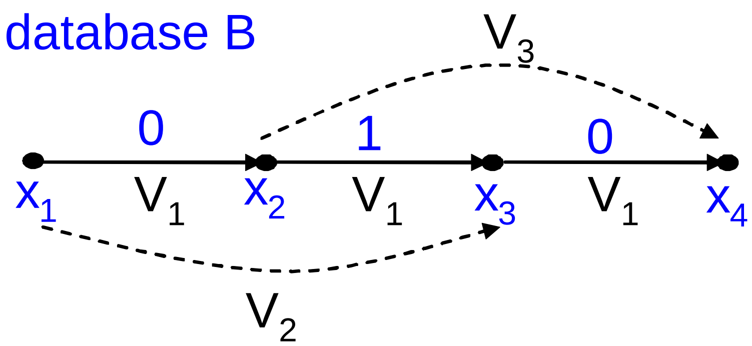
$V_2 = 01$ $V_3 = 10$

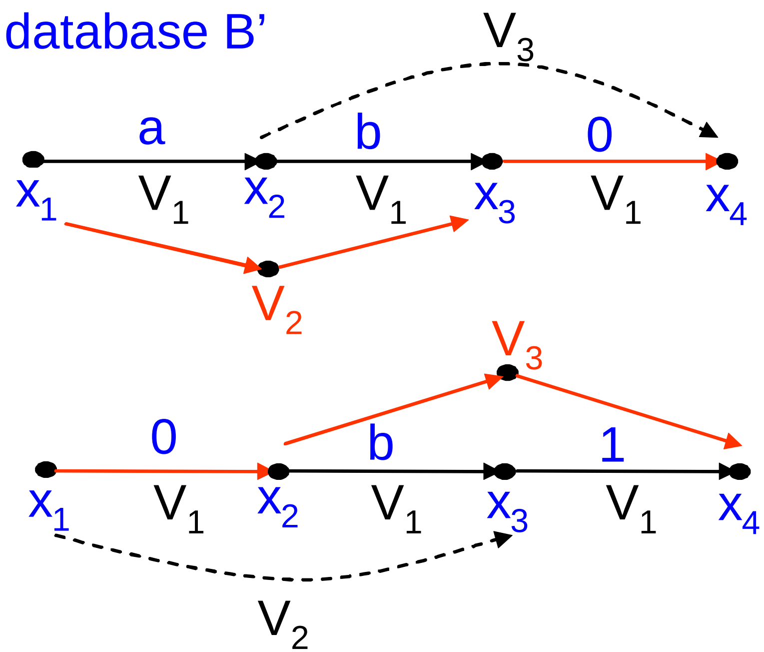
$V_4 = 000$ $V_5 = 111$

Query Q: $010 + 101 + 000 + 111$

$R^{max}_{Q,\mathcal{V}} = V_4 + V_5$

$cert_{Q,\mathcal{V}}$ is equivalent to Q





This shows that losslessness and exactness do not coincide

Loslessness in the case of 2RPQs

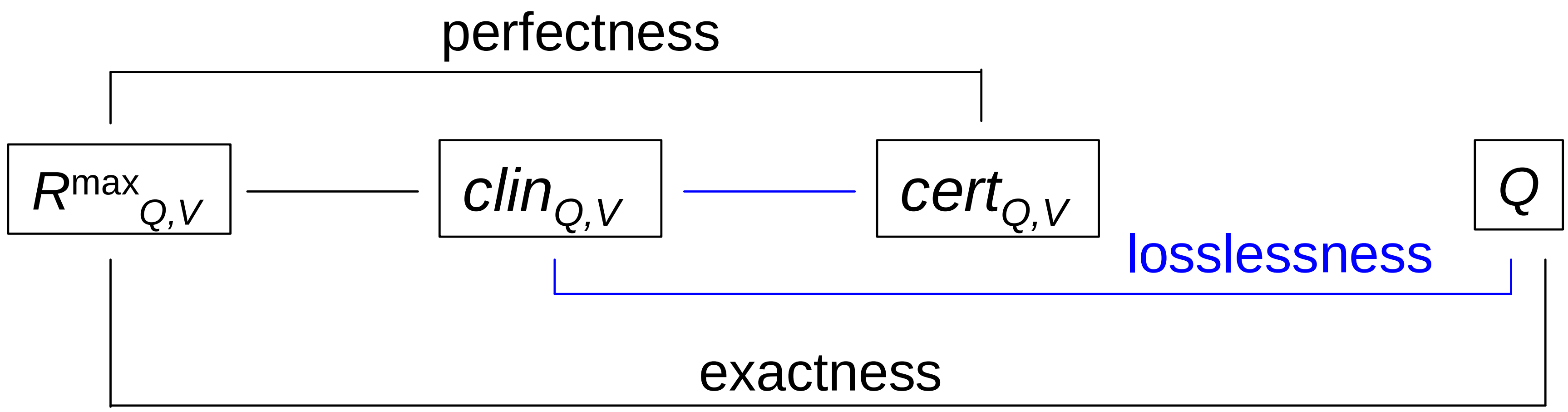
- In [Calvanese&al 2003] we showed that losslessness is EXPSPACE-complete for RPQs
- [CDLV05] shows that losslessness is EXPSPACE-complete also for 2RPQs
- To this end, we introduce the notion of linear approximation to the certain answers

Losslessness in the case of 2RPQs

The linear fragment of certain answers $clin_{Q,\mathcal{V}}$ for a 2RPQ Q wrt a set $\mathcal{V}$ of views is the maximal two-way path query Q' over Σ such that $\forall \mathcal{B}: \; Q'(\mathcal{B}) \subseteq cert_{Q,\mathcal{V}}(\mathcal{V}(\mathcal{B}))$

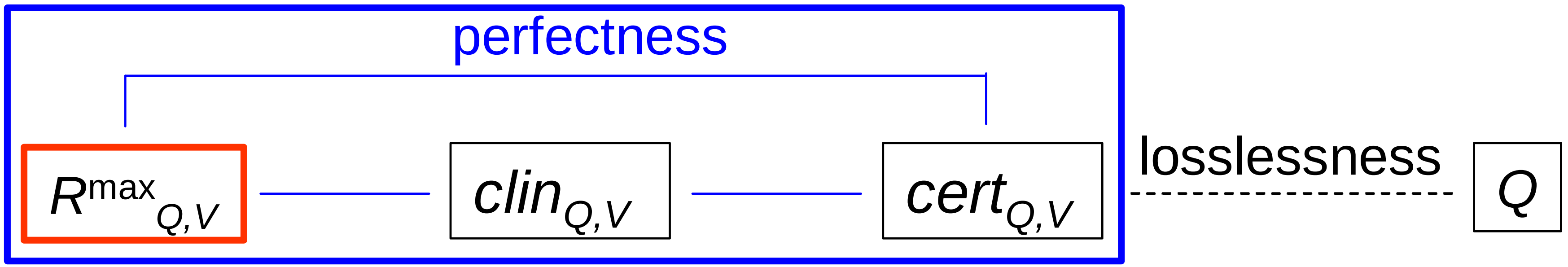
Results in [CDLV05]:

- We have a method for constructing $clin_{Q,\mathcal{V}}$ (always a 2RPQ)
- We show that losslessness means $\forall \mathcal{B} \; Q(\mathcal{B}) \subseteq clin_{Q,\mathcal{V}}(\mathcal{B})$
- Checking losslessness is EXPSPACE-complete



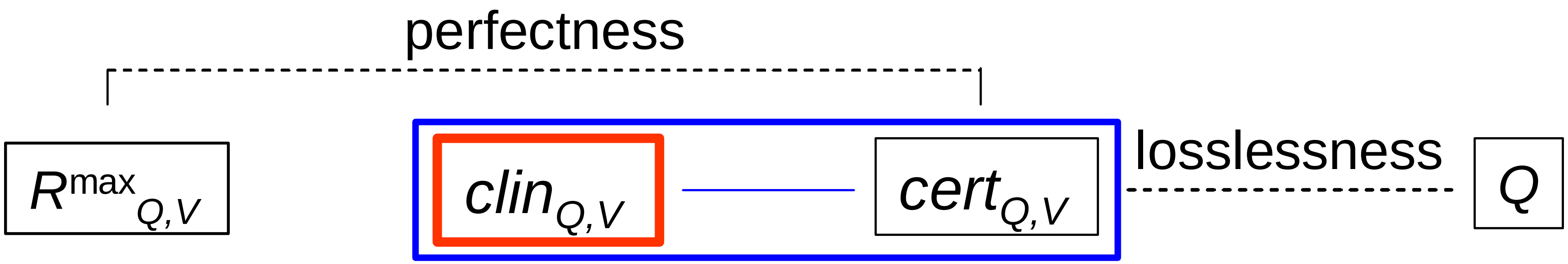
The case of lossiness (with perfectness)

- In case of losslessness, $cert_{Q,\mathcal{V}}$ computes exactly Q (which is equivalent to $clin_{Q,\mathcal{V}}$), and Q explains $cert_{Q,\mathcal{V}}$ at best
- In case of lossiness, still, we would like to express $cert_{Q,\mathcal{V}}$ in the language of the user (2RPQ over the database)
  - If $R^{max}_{Q,\mathcal{V}}$ is perfect, then $R^{max}_{Q,\mathcal{V}}$ is equivalent to $clin_{Q,\mathcal{V}}$, and the unfolding of $R^{max}_{Q,\mathcal{V}}$ explains $cert_{Q,\mathcal{V}}$ at best



The case of lossiness (without perfectness)

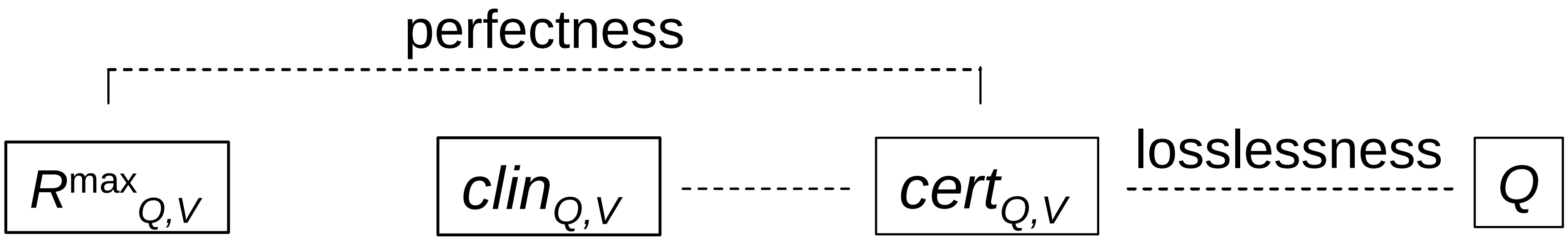
- In case of lossiness, and if $R^{max}_{Q,\mathcal{V}}$ is not perfect, still, we would like to express $cert_{Q,\mathcal{V}}$ as a 2RPQ over the database
  - If $cert_{Q,\mathcal{V}}$ is equivalent to $clin_{Q,\mathcal{V}}$, then $clin_{Q,\mathcal{V}}$ explains $cert_{Q,\mathcal{V}}$ at best (and, if we have access to $\mathcal{B}$, $cert_{Q,\mathcal{V}}$ can be computed by evaluating $clin_{Q,\mathcal{V}}$ over the database)



Result in [CDLV05]: checking whether $cert_{Q,\mathcal{V}}$ is equivalent to $clin_{Q,\mathcal{V}}$ can be done in N3EXPTIME (lower bound open)

The case of lossiness (without perfectness)

- In case of lossiness, and if $R^{max}_{Q,\mathcal{V}}$ is not perfect, and furthermore, if $cert_{Q,\mathcal{V}}$ is not equivalent to $clin_{Q,\mathcal{V}}$, then we would like to exhibit a nonlinear counterexample database explaining lossiness to the user. How to achieve this is an open problem.



Conclusions

- Answering and rewriting are different notions
- Exactness, perfectness and losslessness are different notions
- We introduced the concept of “good” approximation of $cert_{Q,\mathcal{V}}$ for 2RPQs, i.e., the linear fragment
- In our past work, we addressed
  - answering, via the relationship with CSP
  - rewriting, via an automata-theoretic approach

  [CDLV05] also proposes a technique for building $R^{max}_{Q,\mathcal{V}}$ that reconciles the two approaches (see the proceedings)

Future work

- A few lower bounds open
- In the case $cert_{Q,\mathcal{V}}$ is not equivalent to $clin_{Q,\mathcal{V}}$, we would like to exhibit a nonlinear counterexample database explaining lossiness to the user
- Many open problems with exact views:
  - perfectness and losslessness in the case of conjunctive queries and views ($R^{max}_{Q,\mathcal{V}}$ and $cert_{Q,\mathcal{V}}$ do not coincide)
  - perfectness and losslessness in the case of 2RPQs
- More expressive languages, i.e., C2RPQs