Universität Rostock
Traditio et Innovatio

Combining Provenance Management and Schema Evolution

Problem

- Research data management: tracking and archiving of data collected in scientific projects, experiments or observations
- Goal: Traceability, reconstructibility and replicability of the path from data collection to publication

Calculation of a minimal part of the database (minimal sub-database)

- Different constraints for the sub-database to be determined:
  - Number of tuples of the original relation remains unchanged.
  - The sub-database can be mapped homomorphically to the original database.
  - The sub-database is an intensional description of the original database.
- Question: Which additional information is required to be able to reconstruct the minimal part of the database if the result and the evaluation query Q are both archived?
- Idea: Calculation of an inverse query Q_{prov} to determine the minimal sub-database
  $\Rightarrow$ Type of inverse depending on the additional information noted

Unification of Provenance and Evolution

- Goal: Evaluation of provenance queries with changing data and schemas
- Idea: Combination of provenance with schema and data evolution
- Wanted: New minimal sub-database to be archived J^*

$\Rightarrow$ Calculation of a new query $Q'(J(S_3))$ from the old query $Q(I(S_1))$

Provenance Management and Schema Evolution:

- Schemas: S_1, S_2 and S_3
- Minimal sub-databases: $I^* \subseteq I$ and $J^* \subseteq J$
- Input for Q_{prov}: $K^* \subseteq K$
- Query: Q
- Schema evolution: $\mathcal{E}$
- Provenance Query: Q_{prov}

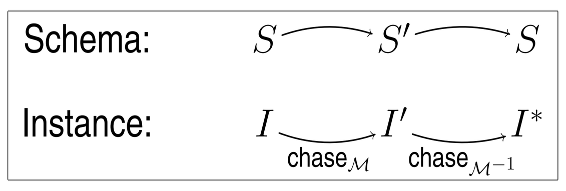
Data Provenance Q_{prov}

- Information order: ***where*** $\preceq$ ***why*** $\preceq$ ***how***
- Provenance types and answers:

Provenance type	Answer
where	tuple or table name
why	(minimal) witness base
how	provenance polynoms
why not	provenance games

CHASE Inverse

- CHASE: Incorporating dependencies $\star$ in an object $\bigcirc$, i.e. $\text{chase}_\star(\bigcirc) = \circledast$
- CHASE&BACKCHASE:



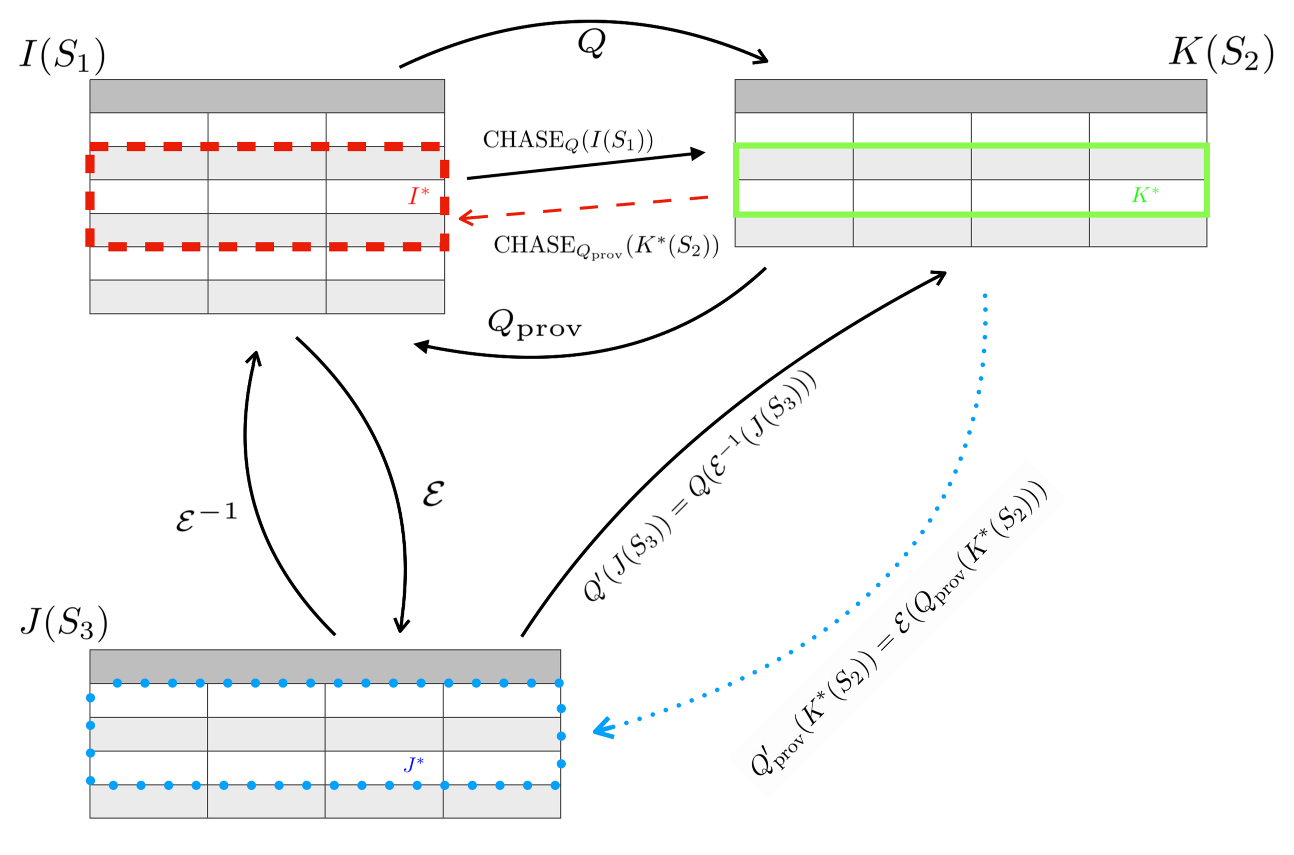
- Exact CHASE-inverse: Reconstructs the complete original database
- Classical CHASE-inverse: Returns a result equivalent to the original database
- Tuple preserving relaxed CHASE-inverse: Preserves the number of tuples
- Result equivalent CHASE-inverse: $\text{chase}_{\mathcal{M}}(I) = \text{chase}_{\mathcal{M}}(I^*)$
- Reduction:

result equivalent $\preceq$ relaxed $\preceq$ tp-relaxed $\preceq$ classical $\preceq$ exact

- Conditions for the existence of CHASE inverse:

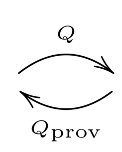
CHASE inverse	sufficient condition	necessary condition
Exact	-	$I^* = I$
Classical	Exact CHASE-inverse	$I^* \equiv I$
Tp-relaxed	Classical CHASE-inverse	$I^* \preceq I, \mid I^* \mid = \mid I \mid$
Relaxed	Tp-relaxed CHASE inverse	$I^* \preceq I$
Result equivalent	Relaxed CHASE-inverse	$I^* \leftrightarrow_{\mathcal{M}} I$

Provenance Management and Schema Evolution:



Example:

id	name	subject	grade	
4711	Tom	C.s.	1.0	t_1
253	Tom	C.s.	2.3	t_2
1933	Tom	Math	1.7	t_3
1234	Peter	Physics	2.0	t_4



name	subject	
Tom	C.s.	$t_1 +_K t_2$
Tom	Math	t_3

id	name	subject	grade	institute	
4711	Tom	C.s.	1.0	Engineering	t_1
253	Tom	C.s.	2.3	Engineering	t_2
1933	Tom	Math	1.7	Science	t_3
1234	Peter	Physics	2.0	Science	t_4
4967	Paul	Math	1.3	Science	t_5

$$
\begin{aligned}
Q &= \pi_{\text{name, subject}}(\sigma_{\text{name = Tom}}(\text{STUDENT}_1)) \\
Q_{\text{prov}} &= (\pi_{\text{name, subject}}(\sigma_{\text{name = Tom}}(\text{STUDENT}_1)))^{-1} \\
&= \sigma^{-1}_{\text{name=Tom}} \circ \pi^{-1}_{\text{name,subject}}(\text{STUDENT}_2)
\end{aligned}
$$

$$
\begin{aligned}
\text{STUDENT}_1(\text{id}, \text{name}, \text{subject}, \text{grade}) &\rightarrow \text{STUDENT}_2(\text{Tom}, \text{subject}) \\
\text{STUDENT}_2(\text{Tom}, \text{subject}) &\rightarrow \exists\, \eta_{\text{id}}, \eta_{\text{grade}} : \text{STUDENT}_1(\eta_{\text{id}}, \text{Tom}, \text{subject}, \eta_{\text{grade}})
\end{aligned}
$$

Query Q

- CHASE algorithm for evaluation of queries
- Approach: Description of the query Q as extended S-T TGDs and EGDs

$\Rightarrow$ Calculation of a CHASE inverse Q_{prod} to reconstruct a minimal sub-database I^*

Schema Evolution $\mathcal{E}$

- CHASE algorithm for schema evolution
- Approach: Description of the schema evolution $\mathcal{E}$ as S-T TGDs and EGDs

$\Rightarrow$ Calculation of an inverse $\mathcal{E}^{-1}$ to reconstruct the old minimal sub-database I^*

Tanja Auge , Andreas Heuer

INSTITUTE OF COMPUTER SCIENCE I DATABASE RESEARCH GROUP I
University of Rostock, Albert-Einstein-Straße 22 I 18059 Rostock, Germany