Composition and Non-Functional Mappings

Alan Nash
Math and CSE Departments
University of California, San Diego

Joint work with
Sergey Melnik and Phil Bernstein

Outline

- Overview
- Motivation: GAV and LAV
- Definitions
- Results
- Composition

Models and Mappings

- A *model* is a set of instances
- A *mapping* is a binary relation on instances

Relational case:

- Instances are databases
- Instances have a *signature*

Composition

Data translation and data exchange:

$$\mathcal{S}_1 \rightarrow \mathcal{S}_2 \rightarrow \mathcal{S}_3 \rightarrow \mathcal{S}_4$$

Global as view:

$$\mathcal{S} \rightarrow \mathcal{G} \rightarrow \mathcal{Q}$$

Local as view, answering queries using views:

$$\mathcal{S} \leftarrow \mathcal{G} \rightarrow \mathcal{Q}$$

Results

- Composition is undecidable
- PTIME conditions for composition
- General algorithm for composition
- Classes of mappings closed under composition
- Other operators on mappings and models

Composition

1. Skolemize $\mathbb{CQ}^{=}$-mappings
2. Find a certain finite axiomatization
3. de-Skolemize the finite axiomatization

Previous Work

- *Composing Mappings Among Data Sources* J. Madhavan and A. Y. Halevy (VLDB 2003)
- *Composing Schema Mappings: Second-Order Dependencies to the Rescue* R. Fagin, P. G. Kolaitis, L. Popa, and W. C. Tan (PODS 2004)

Madhavan and Halevy

- "Language-based" definition of composition
- Examples showing difficulties
- Composition algorithm
- Closure w.r.t. CQ_k queries
- Computing certain answers

Fagin et al.

- “Set-theoretic” definition of composition
- Examples showing difficulties
- Source-to-target limitation
- Second-order dependencies
- Algorithm for composition

Motivation

Data Integration Systems

Adapted from Lenzerini (PODS 2002)

1. Source model $\mathcal{S}$: signature σ + constraints.
2. Global model $\mathcal{G}$: signature γ + constraints.
3. Mapping $m : \mathcal{S} \rightrightarrows \mathcal{G}$: given by constraints.

Note: m is not necessarily *functional*.

Global as View (GAV)

- No constraints on source model.
- Each relation in γ is a view.
- Mapping m is functional.

Example:

- $\sigma := \{A, T\} \quad \gamma := \{AT\}$
- $AT(a, t) = \exists p\, A(a, p), T(p, t)$

Queries in GAV

Give me authors:

$$Aut(a) = \exists t\, AT(a,t)$$

becomes

$$Aut(a) = \exists t, p\, A(a,p), T(p,t)$$

Give me pairs of coauthors:

$$CoA(a,b) = \exists t\, AT(a,t), AT(b,t)$$

becomes $CoA(a,b) =$

$$\exists t, p, p'\, A(a,p), T(p,t), A(b,p'), T(p',t)$$

Local as View (LAV)

- Each relation in σ provides part of a view
- Mapping m is non-functional.

Example:

- $\sigma := \{A, T, VLDB\} \quad \gamma := \{AT\}$
- $\exists t\, AT(a,t) \supseteq A(a)$
- $\exists a\, AT(a,t) \supseteq T(t)$
- $AT(a,t) \supseteq VLDB(a,t)$

For each $A, T, VLDB$, there are many ATs.

Queries in LAV

Give me authors:

$$Aut(a) = \exists t\, AT(a, t)$$

becomes

$$Aut(a) = \exists p A(a, p) \vee \exists t VLDB(a, t)$$

Give me pairs of coauthors:

$$CoA(a, b) = \exists t\, AT(a, t), AT(b, t)$$

becomes

$$CoA(a, b) = \exists t\, VLDB(a, t), VLDB(b, t)$$

Formally, we use *certain answers*.

Mappings and Constraints

A mapping m is given by an expression

$$(\sigma_1, \sigma_2, \Sigma)$$

when

$$(A, B) \in m \text{ iff } (A, B) \models \Sigma$$

where

- A is a database over σ_1
- B is a database over σ_2
- Σ is a set of constraints over $\sigma_1 \cup \sigma_2$

GAV and LAV as Constraints

GAV example:

1. $AT(a,t) \rightarrow \exists p\, A(a,p), T(p,t)$
2. $\exists p\, A(a,p), T(p,t) \rightarrow AT(a,t)$

LAV example:

1. $A(a) \rightarrow \exists t\, AT(a,t)$
2. $T(t) \rightarrow \exists a\, AT(a,t)$
3. $VLDB(a,t) \rightarrow AT(a,t)$

- These are embedded dependencies
- LAV (open world) is source-to-target

Answering Queries

Given a query Q over the global signature γ, the corresponding query over the source signature σ is given by

- $Q_G := m \cdot Q$ in the GAV setting
- $Q_L := \mathrm{certain}(m \cdot Q)$ in the LAV setting

If m is given by CQ queries and $Q \in \mathrm{CQ}$, then

$$Q_G \in \mathrm{CQ} \text{ and } Q_L \in \mathrm{UCQ}.$$

CQ: Conjunctive queries
UCQ: Unions of conjunctive queries

Composition

Given

- $Q'' := m \cdot Q$
- source instance S over σ
- global instance G over γ
- answer instance A over α

$$(S, A) \in m \cdot Q$$

iff

$$\exists G \ (S, G) \in m \text{ and } (G, A) \in Q$$

Certain Answers

Given

- $Q'' := m \cdot Q$
- source instance S over σ
- global instance G over γ
- answer instance A over α

$$(S, A) \in \text{certain}(Q'')$$

iff

$$\forall A' \ (S, A') \in Q'' \rightarrow A \subseteq A'$$

Answering Queries Using Views

Given a view V and a query Q over the source schema we want Q' such that

$$Q := V \cdot Q'$$

That is, we need $Q' := V^{-1} \cdot Q$.
V is functional, but usually V^{-1} is not.
If $V, Q \in \mathrm{CQ}$, then Q' may not be definable in any language due to incomplete information.
This happens precisely when Q' is not functional.

Definitions

Some Query Languages

CQ	Conjunctive queries
UCQ	Unions of conjunctive queries
$\mathrm{CQ}^{=}$	CQ queries with equality (SPJ)
CQ_0	CQ queries without projections
$\mathrm{CQ}_0^{=}$	CQ_0 queries with equality (SJ)

Mappings and Constraints

An $\mathcal{L}$-mapping m is given by $(\sigma_1, \sigma_2, \Sigma)$ when

$$(A, B) \in m \text{ iff } (A, B) \models \Sigma$$

- A is a database over σ_1
- B is a database over σ_2
- Σ is a finite set of constraints over $\sigma_1 \cup \sigma_2$ of the form

$$\{\forall \bar{x}(Q_1(\bar{x}) \rightarrow Q_2(\bar{y})) : Q_1, Q_2 \in \mathcal{L}\}$$

where Q_1, Q_2 are queries in $\mathcal{L}$.

We are interested in the case when $\mathcal{L}$ is one of $\mathrm{CQ}_0, \mathrm{CQ}_0^=, \mathrm{CQ}, \mathrm{CQ}^=$ and some others.
The mappings we have just seen are CQ-mappings

Basic Operators

Models, mappings $\longrightarrow$ models

Domain	$\mathrm{dom}(m)$
Range	$\mathrm{rng}(m)$
Intersection	$\mathcal{A} \cap \mathcal{B}$

Models, mappings $\longrightarrow$ mappings

Identity	$\mathrm{id}(\mathcal{A})$
Cross product	$\mathcal{A} \times \mathcal{B}$
Intersection	$m_1 \cap m_2$
Composition	$m_1 \cdot m_2$
Inverse	m^{-1}

Operators: Definitions

$$
\begin{aligned}
\mathrm{dom}(m) &:= \{A : \exists B \, \langle A, B\rangle \in m\}. \\
\mathrm{rng}(m) &:= \{B : \exists A \, \langle A, B\rangle \in m\}. \\
\mathcal{A} \cap \mathcal{B} &:= \{A : A \in \mathcal{A}, A \in \mathcal{B}\}. \\
\mathrm{id}(\mathcal{A}) &:= \{\langle A, A\rangle : A \in \mathcal{A}\}. \\
\mathcal{A} \times \mathcal{B} &:= \{\langle A, B\rangle : A \in \mathcal{A}, B \in \mathcal{B}\}. \\
m_1 \cap m_2 &:= \{\langle A, B\rangle : \langle A, B\rangle \in m_1, \langle A, B\rangle \in m_2\}. \\
m_1 \cdot m_2 &:= \{\langle A, C\rangle : \exists B \, \langle A, B\rangle \in m_1, \langle B, C\rangle \in m \\
m^{-1} &:= \{\langle B, A\rangle : \langle A, B\rangle \in m\}.
\end{aligned}
$$

Results

Basic Questions

Given an operator $\mathcal{O}$ and $\mathcal{L}$-models and $\mathcal{L}$-mappings for input:

1. Is the output always an $\mathcal{L}$-model or an $\mathcal{L}$-mapping? (In this case we say that $\mathcal{L}$ is *closed* under $\mathcal{O}$.)
2. If not, is there a decision procedure to determine when the output is an $\mathcal{L}$-model or an $\mathcal{L}$-mapping?

Basic Results

Proposition 1. *Every $\mathcal{L} \supseteq \mathrm{CQ}_0$ is closed under identity, cross product and intersection.*

Proposition 2. *Each one of the operators composition, range, and domain can be reduced to any one of the others.*

Composition and Closure

Consider CQ_0-mappings m_{12} and m_{23} given by

$$\begin{array}{lrcl} \Sigma_{12}: & R(xy) & \rightarrow & S(xy) \\ & S(xy), S(yz) & \rightarrow & S(xz) \\ \Sigma_{23}: & S(xy) & \rightarrow & T(xy) \end{array}$$

where $\sigma_1 = \{R\}$, $\sigma_2 = \{S\}$, and $\sigma_3 = \{T\}$.
These say:

$$R \subseteq S = \mathrm{tc}(S) \subseteq T$$

and this implies $\mathrm{tc}(R) \subseteq T$.
$m_{12} \cdot m_{23}$ is not expressible in FO.

Undecidability

Theorem 1. *Checking whether the composition of two* CQ_0*-mapping is a* CQ_0*-mappings is undecidable (in fact, coRE-hard). The same holds with* $\mathrm{CQ}_0^{=}$ *instead of* CQ_0.

Corollary 1. *Checking whether the domain or range of a* CQ_0*-mapping is a* CQ_0*-model is undecidable.* The same holds for CQ and $\mathrm{CQ}^{=}$.

Summary of Other Results

- We provide necessary and sufficient conditions for the composition of two $\mathrm{CQ}_0^=$-mappings to be a $\mathrm{CQ}_0^=$-mapping (same for $\mathrm{CQ}^=$).
- We provide sufficient conditions which can be checked in polynomial time.
- We provide a general algorithm for computing the composition when these conditions are satisfied.
- We define subclasses of $\mathrm{CQ}_0^=$ and $\mathrm{CQ}^=$ which are closed under composition.

Composition

Main Theorem

Theorem 2. *If the $\mathrm{CQ}_0^=$-mappings m_1, m_2 are given by $(\sigma_1, \sigma_2, \Sigma_{12})$ and $(\sigma_2, \sigma_3, \Sigma_{23})$ with $\sigma_{13} = \sigma_1 \cup \sigma_3$ and $\Sigma_{123} := \Sigma_{12} \cup \Sigma_{23}$, then these are equivalent*

1. *There is a finite set of constraints $\Sigma_{13} \subseteq \mathrm{IC}(\mathrm{CQ}_0^=)$ over the signature σ_{13} such that $m := m_1 \cdot m_2$ is given by $(\sigma_1, \sigma_3, \Sigma_{13})$.*
2. *There is a finite set of constraints $\Sigma_{13} \subseteq \mathrm{IC}(\mathrm{CQ}_0^=)$ over the signature σ_{13} such that*

$$\mathrm{DC}(\mathrm{CQ}_0^=, \Sigma_{123})|_{\sigma_{13}} = \mathrm{DC}(\mathrm{CQ}_0^=, \Sigma_{13}).$$

3. *There is m such that for every ξ over σ_{13} satisfying $\Sigma_{123} \vdash \xi$ there is a deduction of ξ from Σ_{123} using at most m σ_2-resolutions.*

Composition Algorithm

Algorithm Compose(Σ_{12}, Σ_{23}):
Set $\Sigma := \Sigma_{12} \cup \Sigma_{23}$
Repeat
  Set $\Sigma' := \emptyset$
  For every pair $\phi, \psi \in \Sigma$
    If ϕ, ψ can be σ_2-unified to yield ξ
      and there is no variant of ξ in Σ
      then set $\Sigma' := \Sigma' \cup \{\xi\}$
  Set $\Sigma := \Sigma \cup \Sigma'$
Until $\Sigma' = \emptyset$
Return $\Sigma_{13} := \Sigma|\sigma_{13}$

Correctness

Corollary 2. *Under the hypotheses of Theorem 2, Compose(Σ_{12}, Σ_{23}), whenever it terminates, yields Σ_{13} such that $m_{12} \cdot m_{23}$ is given by $(\sigma_1, \sigma_3, \Sigma_{13})$.*

Composition Example

Consider CQ_0-mappings m_{12} and m_{23} given by

$$\begin{array}{lrcl} \Sigma_{12}: & R(xy) & \rightarrow & S(xy) \\ & S(xy), S(yz) & \rightarrow & R(xz) \\ \Sigma_{23}: & S(xy) & \rightarrow & T(xy) \end{array}$$

where $\sigma_1 = \{R\}$, $\sigma_2 = \{S\}$, and $\sigma_3 = \{T\}$.

The constraints

$$\begin{array}{rcl} R(xy), R(yz) & \rightarrow & R(xz) \\ R(xy) & \rightarrow & T(xy) \end{array}$$

express exactly the composition $m_1 \cdot m_2$, and are exactly those found by Compose(Σ_{12}, Σ_{23}).

Checking for Composability

Theorem 3. *Under the hypotheses of Theorem 2, if no constraint of the form* $\phi(\bar{z}), S(\bar{y}) \rightarrow S(\bar{x})$ *where*

1. $\phi(\bar{z})$ *is a conjunction of atoms over* σ_{123},
2. $\{\bar{x}\} \not\subseteq \{\bar{y}\}$, *and*
3. S *is a relation symbol in* σ_2

(we call this a bad *constraint) can be deduced from* Σ_{123} *using only* σ_2*-unifications, then Compose*$(\Sigma_{12}, \Sigma_{23})$ *terminates and therefore* $m_1 \cdot m_2$ *is a* $\mathrm{CQ}_0^=$*-mapping.*
The same holds for CQ_0*-mappings.*

Closure under Composition

Definition 1. A $\mathrm{CQ}_0^=$-mapping is a *good-*$\mathrm{CQ}_0^=$*-mapping* if it is given by $(\sigma_1, \sigma_2, \Sigma_{12})$ such that no constraint of the form $\phi(\bar{z}), S(\bar{y}) \rightarrow S'(\bar{x})$ where

1. $\phi(\bar{z})$ is a conjunction of atoms over $\sigma_1 \cup \sigma_2$,
2. $\{\bar{x}\} \not\subseteq \{\bar{y}\}$, and
3. S and S' are both relation symbol in σ_1 or both in σ_2

can be deduced from Σ_{123} using only σ_1-unifications or only σ_2-unifications. We define good-CQ_0 similarly.

Theorem 4. *good-*$\mathrm{CQ}_0^=$ *and good-*CQ_0 *are closed under composition and inverse.*

Skolem Functions

Skolem functions replace $\exists$-quantifed variables.
For example, $AT(a,t) \rightarrow \exists p\, A(a,p), T(p,t)$
becomes $AT(a,t) \rightarrow A(a, f(a,t)), T(f(a,t),t)$.

Composition (GAV)

1. $AT(a,t) \rightarrow A(a, f(a,t)), T(f(a,t),t)$
2. $A(a,p), T(p,t) \rightarrow AT(a,t)$
3. $CoA(a,b) \rightarrow AT(a, g(a,b)), AT(b, g(a,b))$
4. $AT(a,t), AT(b,t) \rightarrow CoA(a,b)$

γ-unification gives:

- $A(a,p), T(p,t), A(b,p'), T(p',t) \rightarrow CoA(a,b)$ from 2,4
- $CoA(a,b) \rightarrow$
  $A(a, f(a, g(a,b))), T(f(a, g(a,b)), t)$
  $A(b, f(b, g(a,b))), T(f(b, g(a,b)), t)$ from 1,3

$CoA(a,b) \rightarrow \exists t,p,p'\, A(a,p), T(p,t), A(b,p'), T(p',t)$

Composition (LAV)

1. $A(a) \rightarrow AT(a, f(a))$
2. $T(t) \rightarrow AT(h(t), t)$
3. $VLDB(a, t) \rightarrow AT(a, t)$
4. $CoA(a, b) \rightarrow AT(a, g(a, b)), AT(b, g(a, b))$
5. $AT(a, t), AT(b, t) \rightarrow CoA(a, b)$

γ-unification gives:

- $A(a), A(b), f(a) = f(b) \rightarrow$
  $CoA(a, f(a)), CoA(b, f(b))$
- $A(a), T(t), f(a) = t \rightarrow CoA(a, t), CoA(h(t), t))$
- $VLDB(a, t), VLDB(b, t) \rightarrow CoA(a, b)$

$$VLDB(a, t), VLDB(b, t) \leftrightarrow CoA(a, b)$$

De-Skolemization

Example from Fagin et al. (PODS 2004)

$$\begin{array}{lrcl} \Sigma_{12}: & E(x,y) & \rightarrow & F(xy) \\ & E(x,y) & \rightarrow & \exists u\, C(x,u) \\ \Sigma_{23}: & C(x,u), C(y,v), F(x,y) & \rightarrow & D(u,v) \end{array}$$

where $\sigma_1 = \{E\}$, $\sigma_2 = \{F, C\}$, and $\sigma_3 = \{D\}$.
We deduce: $C(x, g(x,y)), C(y, g(y,v)), F(x,y)$ $\rightarrow D(g(x,y), g(y,v))$
If $D := \{(r,g),(r,b),(g,r),(g,b),(b,r),(b,g)\}$ and E is undirected then

$$(E, D) \in m_{12} \cdot m_{23} \text{ iff } E \text{ is 3-colorable}$$

Therefore $m_{12} \cdot m_{23}$ is not a $\mathrm{CQ}^{=}$-mapping.