Determining Source Contribution in Integration Systems

A presentation by
Yannis Katsis

on a joint work with
Alin Deutsch
Yannis Papakonstantinou
The Problem

Client Queries

Mediator
(Global DB)

Sources (Actual Local DBs)
The Problem

- What is the contribution of source $S$ to the result of the query $Q$?
The Problem

- What is the contribution of source $S$ to the result of the query $Q$?

$Q$: `cars`  
- $S$ is Self Sufficient w.r.t. $Q$

$Q$: `cars JOIN reviews`  
- $S$ is Now Complementary w.r.t. $Q$
# Roadmap

## Framework

<table>
<thead>
<tr>
<th>Degrees of Contribution</th>
</tr>
</thead>
</table>

| Decidability Results    |
Relational Schemas: Local and Global

- Relational Schemas
- Visual Representation

Source 1: Business Magazine

Source 2: Car Magazine

Global: Car Portal

Source 1: Business Magazine
Source 2: Car Magazine
Global: Car Portal
Source Registration using GLAV Mappings

- **Source Registration:**
  Correspondence between a source schema and the global schema

  =

  Set of Mapping Constraints of the form

  \[(U \subseteq V)\]

  CQ= over source schema

  CQ= over global schema

- Open World
- Global and Local As View (GLAV)
Target Constraints

• Constraints on the global schema
  =
  Set of Constraints of the form

\( (U \subseteq V) \)

\( CQ^= \) over global schema  \( CQ^= \) over global schema

• Expresses Embedded Dependencies (PKs, RICs, ...)
Visual Representation of Mappings (1)

Business Magazine: Provides Carmake and Origin

\[ (U_1 \subseteq V_1) \]

\[ U_1(C, O) :- \text{make}(C, O, S) \]

\[ V_1(C, O) :- \text{brand}(C, O) \]
Visual Representation of Mappings (2)

• Visual Representation (IBM Clio)

Car Magazine: Provides Model, Carmake and Baseprice
Query Semantics

- Queries in UCQ=
- Set of Possible Global Instances
  *Set of global instances that satisfy all constraints*
- Query Answers = Set of Certain Answers
  *The tuples appearing in the answer to Q for any possible global instance*
Roadmap

- Framework
- Degrees of Contribution
- Decidability Results
Source Instance’s Contribution

- For given instances of the sources

Contribution to Q of Source Instance

\[ \text{Contribution to } Q \text{ of Source Instance} = \text{The tuples in answer of } Q \text{ not provided by the other sources} \]
Source Registration’s Contribution

- Source Registration: Source Mappings
- Degrees of Source Registration’s Contribution

1. Self Sufficient
2. Now Complementary
3. Later Complementary
4. Unusable

More contribution
Less contribution
Self Sufficient Registration: Example

**Example**

- **Baseprices of Models**

<table>
<thead>
<tr>
<th>car</th>
<th>Model</th>
<th>Carmake</th>
<th>Doors</th>
<th>Baseprice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M3</td>
<td>BMW</td>
<td></td>
<td>45K</td>
</tr>
</tbody>
</table>

\[ G \]
- **car**
  - Model
  - Carmake
  - Doors
  - Baseprice
- **brand**
  - Carmake
  - Origin

*Green Registration is Self Sufficient*
Self Sufficient Registration: Definition

Self Sufficient

∃ Source instance s.t.

The source has a non empty contribution in the absence of the other sources

Answer to Q ≠ ∅
Now Complementary Registration: Example

Baseprices of Models by German manufacturers

Example

\[ G \]

- car
  - Model
  - Carmake
  - Doors
  - Baseprice

- brand
  - Carmake
  - Origin = 'Germany'

Green Registration is Now Complementary
Now Complementary Registration: Definition

2. Now Complementary

Not Self Sufficient

\&

∃ Source instances

s.t.

The source has a non empty contribution in combination with the other existing sources

Answer to Q \neq \emptyset

Answer to Q
Later Complementary Registration: Example

Baseprices of Models by German manufacturers

<table>
<thead>
<tr>
<th>car</th>
<th>brand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Carmake</td>
</tr>
<tr>
<td>Doors</td>
<td>Origin = 'Germany'</td>
</tr>
<tr>
<td>Baseprice</td>
<td></td>
</tr>
</tbody>
</table>

Example:

- car
  - Model: M3
  - Doors: ?
  - Baseprice: 45K

- brand
  - Carmake: BMW
  - Origin: Germany

Green Registration is Later Complementary
Later Complementary Registration: Definition

3 Later Complementary

Not Self Sufficient &
Not Now Complementary
&
∃ Potential future sources
& Source instances

s.t.
The source has a non empty contribution in combination with the future sources

Answer to Q

Answer to Q

≠ ∅
Green Registration is Unusable
4 Unusable

Not Self Sufficient & Not Now Complementary & Not Later Complementary

⇔

The source has a empty contribution regardless of what sources enter the system

Answer to Q

Answer to Q

= ∅
Roadmap

- Framework
- Degrees of Contribution
- Decidability Results
Subtleties for Unusable Registrations

Example

Baseprices and Doors of Models

<table>
<thead>
<tr>
<th>car</th>
<th>Model</th>
<th>Carmake</th>
<th>Doors</th>
<th>Baseprice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M3</td>
<td>BMW</td>
<td>2</td>
<td>45K</td>
</tr>
</tbody>
</table>

Green Registration is Unusable
In presence of PK Unusable Example becomes Later Complementary

Example

Baseprices and Doors of Models

<table>
<thead>
<tr>
<th>car</th>
<th>Model</th>
<th>Carmake</th>
<th>Doors</th>
<th>Baseprice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M3</td>
<td>BMW</td>
<td>2</td>
<td>45K</td>
</tr>
</tbody>
</table>

Green Registration is Later Complementary
## Decidability Results

### Overview: What is decidable

<table>
<thead>
<tr>
<th>Degrees</th>
<th>Self Sufficient</th>
<th>Now complementary</th>
<th>Later complementary</th>
<th>Unusable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target constraints</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Primary keys</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Primary keys + Referential Integrity Constraints</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Certain Rewriting

Original Query
Query Q over global schema

Certain Rewriting
Query rewr(Q) over source schemas

Certain Answers to Q = Answers to rewr(Q)
Decidability of Self Sufficiency

Self Sufficiency

- **Reduction**: Self Sufficiency \(\Rightarrow\) Satisfiability of certain rewriting

- **PKs**: Certain rewriting expressible as a potentially recursive datalog program [Duschka, Genesereth]
  
  \[\text{Self Sufficiency decidable}\]

- **PKs & RICs**: No known result for finding certain rewriting

  Reduction from the Post Correspondence Problem [new result]

  \[\text{Self Sufficiency undecidable}\]
Decidability of Now Complementarity

Now Complementarity

- **Reduction**: Now Complementarity \(\rightarrow\) Containment of certain rewritings

- **PKs**: Certain rewriting expressible as a potentially recursive datalog program [Duschka, Genesereth]

  Certain rewriting expressible as UCQ= [new result]

  \(\rightarrow\) Now Complementarity decidable

- **PKs & RICs**: Reduction from Self Sufficiency

  \(\rightarrow\) Now Complementarity undecidable
Decidability of Later Complementarity

Later Complementarity

- **No target constraints:** Check whether source registration is now complementary with identity registration

<table>
<thead>
<tr>
<th>None</th>
<th>PKs</th>
<th>PKs &amp; RICs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decidable</td>
<td>Decidable</td>
<td>?</td>
</tr>
</tbody>
</table>

- **PKs:** This approach does not work. Check if now complementary with more registrations
• Summary:

- Definition of 4 Degrees of Contribution
- Decidability results based on the type of target constraints
- Extension of work on Certain Rewritings

• Future Work:

- Make definitions instance dependent
- Extend classes of supported queries and mappings
- Interface for guiding the source owner through source registration