Universal Models for Guarded Tgds

André Hernich
Universal Models
Data Exchange
(Fagin, Kolaitis, Miller, Popa '03)

Goal: **Materialize** a target database.
Specifying the Translation

- Source schema $S$
  - Amtrak(from, to, fare)
  - Caltrain(from, to)

- Target schema $T$
  - Conn(from, to, fare)

- Set $\Sigma$ of declarative assertions / constraints
  
  Amtrak($x, y, z$) $\rightarrow$ Conn($x, y, z$)
  Caltrain($x, y$) $\rightarrow$ $\exists z$ Conn($x, y, z$)
  Conn($x, y, z$) $\land$ Conn($x, y, z'$) $\rightarrow$ $z = z'$
Types of Constraints 1/2

**Tuple-Generating Dependencies (tgds)**

\[ R_1(x_1) \land \cdots \land R_k(x_k) \rightarrow \exists z \; S_1(y_1) \land \cdots \land S_l(y_l) \]

contains only variables from the body and \( z \)

**Example:**

\[ \text{Amtrak}(x, y, u) \land \text{Caltrain}(y, z) \rightarrow \exists v \; \text{Conn}(x, z, v) \]
Types of Constraints 2/2

**Equality-Generating Dependencies (egds)**

\[ R_1(\bar{x}_1) \land \cdots \land R_k(\bar{x}_k) \rightarrow y = z \]

**Example:**

\[ \text{Conn}(x, y, z) \land \text{Conn}(x, y, z') \rightarrow z = z' \]
Admissible Target Databases

**Amtrak**
- from, to, fare

**Caltrain**
- from, to

**Conn**
- from, to, fare

### Solutions

<table>
<thead>
<tr>
<th>from</th>
<th>to</th>
<th>fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santa Cruz</td>
<td>San Jose</td>
<td>$5</td>
</tr>
<tr>
<td>San Jose</td>
<td>Palo Alto</td>
<td>X</td>
</tr>
</tbody>
</table>

### Train Data

<table>
<thead>
<tr>
<th>from</th>
<th>to</th>
<th>fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santa Cruz</td>
<td>San Jose</td>
<td>$5</td>
</tr>
<tr>
<td>San Jose</td>
<td>Palo Alto</td>
<td>X</td>
</tr>
<tr>
<td>San Jose</td>
<td>Gilroy</td>
<td>$10</td>
</tr>
<tr>
<td>San Jose</td>
<td>Palo Alto</td>
<td>$42</td>
</tr>
</tbody>
</table>
Universal Solutions
(Fagin, Kolaitis, Miller, Popa '03)

Universal solution:
- a solution \( J \)
- for every other solution \( J' \) there is a homomorphism from \( J \) to \( J' \)
Example Revisited

Amtrak(from, to, fare)
Caltrain(from, to)
Conn(from, to, fare)

Amtrak\( (x, y, z) \rightarrow \text{Conn}(x, y, z) \)
Caltrain\( (x, y) \rightarrow \exists z \text{Conn}(x, y, z) \)
Conn\( (x, y, z) \land \text{Conn}(x, y, z') \rightarrow z = z' \)

<table>
<thead>
<tr>
<th>from</th>
<th>to</th>
<th>fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santa Cruz</td>
<td>San Jose</td>
<td>$5</td>
</tr>
<tr>
<td>San Jose</td>
<td>Palo Alto</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>from</th>
<th>to</th>
<th>fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santa Cruz</td>
<td>San Jose</td>
<td>$5</td>
</tr>
<tr>
<td>San Jose</td>
<td>Palo Alto</td>
<td>X</td>
</tr>
<tr>
<td>San Jose</td>
<td>Gilroy</td>
<td>$10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>from</th>
<th>to</th>
<th>fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santa Cruz</td>
<td>San Jose</td>
<td>$5</td>
</tr>
<tr>
<td>San Jose</td>
<td>Palo Alto</td>
<td>$42</td>
</tr>
</tbody>
</table>
Using Universal Solutions

**Theorem** (Fagin, Kolaitis, Miller, Popa '03)
If $I$ is a source database, and $Q$ is a Boolean conjunctive query over the target schema:

$$Q \text{ is true in all solutions for } I \iff Q \text{ is true in a universal solution for } I$$
Further Applications of Universal Solutions

- Data integration
- Answering conjunctive queries over incomplete databases
- ...

Answering Conjunctive Queries over Incomplete Databases

Given: Database D, constraints Σ (tgds & egds), Boolean conjunctive query Q
Task: Decide whether Q is true in all models of D and Σ.

- **Model of D and Σ**: (possibly infinite) database containing D and satisfying Σ
- **Universal model**: finite model with homomorphisms into all other models
How to Compute Universal Models?
Undecidability

**Theorem (H., Schweikardt '07)**
Existence of universal models is undecidable, even for some fixed set of tgdss. 
The Chase (Beeri, Vardi '84)

\[ \Sigma: \ R(x, y) \land P(x) \rightarrow \exists z \ S(y, z) \quad \text{D = \{ R(a,b), P(a) \}} \]

\[ S(y, z) \rightarrow P(z) \]

\[ S(y, z) \rightarrow \exists u \ R(z, u) \]

Chase:
1) \{ R(a,b), P(a) \}
2) \{ R(a,b), P(a), S(b,X) \}
3) \{ R(a,b), P(a), S(b,X), P(X) \}
4) \{ R(a,b), P(a), S(b,X), P(X), R(X,Y) \}
5) \{ R(a,b), P(a), S(b,X), P(X), R(X,Y), S(Y,Z) \}
6) ...

Result of the chase = union of all these databases
Basic Properties

- If the chase terminates, then its result is a universal model (Fagin, Kolaitis, Miller, Popa '03).
- The chase may not terminate. Termination is undecidable.
Termination Conditions

- Several *sufficient* conditions for chase termination are known.

- But they don't cover sets of basic database constraints like
  - inclusion dependencies
  - foreign key constraints
Guarded Tgds?
Guarded Tgds
(Calì, Gottlob, Kifer '08)

Tgds of the following form:

\[
R_1(\bar{x}_1) \land \cdots \land R_k(\bar{x}_k) \rightarrow \exists \bar{z} \ S_1(\bar{y}_1) \land \cdots \land S_l(\bar{y}_l)
\]

Has one atom containing all the variables in the body.

Examples:

- \[ R(y, z, x) \land E(x, y) \land P(z) \rightarrow \exists u \ R(x, z, u) \land P(u) \]
- all inclusion dependencies
Query Answering under Guarded Tgds

Theorem (Calì, Gottlob, Kifer '08)

Let $\Sigma$ be a set of guarded tgds and $Q$ a Boolean conjunctive query.

Testing whether $Q$ is true in all models of a given database and $\Sigma$ is in PTIME.
Main Result

**Theorem** (H., 2012)

Let $\Sigma$ be a set of guarded tgds. The following problem is in PTIME:

Input: Database $D$

Task: Decide whether there is a universal model for $D$ and $\Sigma$. If so, compute one.
Guarded Chase Forests
(Calì, Gottlob, Kifer '08)

\[ \Sigma: \quad R(x, y) \land P(x) \rightarrow \exists z \ S(y, z) \]

\[ S(y, z) \rightarrow P(z) \]

\[ S(y, z) \rightarrow \exists u \ R(z, u) \]

\( D = \{ \, R(a,b), \, P(a) \, \} \)
Main Lemma

There is a number $k$ (depends only on $\Sigma$) s.t.

- There is a universal model of $D$ and $\Sigma$ iff
- The first $k$ levels of the guarded chase forest contain such a model.
Proof Overview

Guarded chase forest for $D$ and $\Sigma$

Pieces of the universal model

Smallest subforest enclosing the model
Proof Overview

Nodes belonging to the universal model

Guarded chase forest for $D$ and $\Sigma$

Step 1: Bound number of red/black nodes on each path.
Step 2: Bound number of nodes “in between”.

Branch nodes
Summary

• Universal models under guarded tgds can be computed in polynomial time (data complexity)
  – We may add certain key constraints, and negative constraints
  – Extension to input databases with nulls
  – Extends to weak universal models (not with keys)

• Open:
  – Combined complexity
  – Weak universal models under guarded tgds + keys
  – Better algorithms for special cases