Computing Range Consistent Answers to Aggregation Queries via Rewriting [1]

Aziz Amezian El Khalfioui and Jef Wijsen

University of Mons, Belgium

November 22, 2024

[1] This paper will also be presented at PODS 2025 [AEKW24].

Table of Contents

Numerical queries

We consider numerical queries that take the following form

$$\mathtt{AGG}(r) \leftarrow q(\vec{u})$$

where

- $\mathtt{AGG}$ is an aggregate symbol (like $\mathtt{MAX}$, $\mathtt{MIN}$, $\mathtt{SUM}$, $\mathtt{AVG}$, $\mathtt{COUNT}$);
- $q(\vec{u})$ is a self-join-free conjunction of atoms with variables $\vec{u}$; and
- r is either a numeric variable occurring in $\vec{u}$ or a constant rational number.

Definition

Let **db** be a database instance. An embedding of $q(\vec{u})$ in **db** is a total mapping θ from $\vec{u}$ to the set of constants appearing in **db** such that $\theta(q) \subseteq \mathbf{db}$.

Semantics of Numerical Queries

Example

Consider the conjunction $Dealers(\underline{x}, y) \land Stock(\underline{z, y}, r)$, and the following database instance **db**:

Dealers

Name	*Town*
Smith	Boston
Smith	New York

Stock

Product	*Town*	*Qty*
Tesla X	Boston	35
Tesla Y	Boston	35
Tesla Y	New York	90

We have 3 different embeddings in **db**:

- one for Tesla X sold in Boston;
- one for Tesla Y sold in Boston; and
- one for Tesla Y sold in New York.

Semantics of Numerical Queries

Example

Consider the numerical query

$$g() = \texttt{SUM}(r) \leftarrow \mathit{Dealers}(\underline{x}, y) \wedge \mathit{Stock}(\underline{z, y}, r)$$

and the following database instance **db**:

Dealers

Name	*Town*
Smith	Boston
Smith	New York

Stock

Product	*Town*	*Qty*
Tesla X	Boston	35
Tesla Y	Boston	35
Tesla Y	New York	90

- $g()$ returns the sum over the multiset containing $\theta(r)$ for each embedding θ in **db**.
- In this example, $g()$ returns $\texttt{SUM}(\{\{35, 35, 90\}\}) = 160$.

Table of Contents

Inconsistent Databases

Primary Key Violations

We consider database instances that may violate primary key constraints.

Example

Consider the following database instance:

Dealers

Name	*Town*
Smith	Boston
Smith	New York

Stock

Product	*Town*	*Qty*
Tesla X	Boston	35
Tesla Y	Boston	35
Tesla Y	New York	90

The primary key *Name* $\rightarrow$ *Town* is violated.

Inconsistent Databases

Definition

Let **db** be a database instance. A repair is an inclusion-maximal subset that satisfies all primary keys.

Example

Consider the following database instance:

Dealers

Name	*Town*
Smith	Boston
Smith	New York

Stock

Product	*Town*	*Qty*
Tesla X	Boston	35
Tesla Y	Boston	35
Tesla Y	New York	90

We have two repairs:

- one where the dealer Smith works in Boston; and
- one where the dealer Smith works in New York.

Table of Contents

Range Consistent Answers

We employ the range semantics as presented by Arenas et al. [ABC01], which provides the greatest lower bound (glb) and the least upper bound (lub) of query answers across all repairs.

Example

Consider the numerical query $\mathtt{SUM}(r) \leftarrow \mathit{Dealers}(\underline{x}, y) \wedge \mathit{Stock}(\underline{z, y}, r)$ and the following database instance:

Dealers

<u>*Name*</u>	*Town*
Smith	Boston
Smith	New York

Stock

<u>*Product*</u>	<u>*Town*</u>	*Qty*
Tesla X	Boston	35
Tesla Y	Boston	35
Tesla Y	New York	90

- The numerical query returns $\mathtt{SUM}(\{\{35, 35\}\}) = 70$ on one repair, and $\mathtt{SUM}(\{\{90\}\}) = 90$ on the other repair.
- Thus, the glb is 70, the lub is 90, and the range consistent answer is $[70, 90]$.

Range Consistent Answers

Definition

Let $g()$ be a numerical query. The function problems GLB-CQA($g()$) and LUB-CQA($g()$) take a database instance **db** as input, and return, respectively, the glb and the lub of the set that contains each number returned by $g()$ on some repair.

In this work, we focus on GLB-CQA($g()$).

Table of Contents

Main Result

Definition

An aggregate operator is associative if for all non-empty multisets X and Y such that $X \neq \emptyset$, we have $\mathcal{F}(X \uplus Y) = \mathcal{F}(\{\{\mathcal{F}(X)\}\} \uplus Y)$, where $\uplus$ denotes union of multisets.

Definition

An aggregate operator $\mathcal{F}$ is monotone if for all $m > 0$ and every (possibly empty) multiset Y, we have $\mathcal{F}(\{\{x_1, \ldots, x_m\}\}) \leq \mathcal{F}(\{\{x'_1, \ldots, x'_m\}\} \uplus Y)$ whenever $x_i \leq x'_i$ for every i.

Example

`MAX` and `SUM` (over non-negative values) are monotone and associative.
`MIN` is associative but not monotone since $\texttt{MIN}(\{\{2, 3\}\}) \not\leq \texttt{MIN}(\{\{2, 3, 1\}\})$.

Main Result

The logic AGGR[FOL] extends FOL with aggregation operators.

Theorem

The following decision problem is decidable in quadratic time (in the size of the input):

> *Given a numerical query* $g() := \mathtt{AGG}(r) \leftarrow q(\vec{u})$ *such that the aggregate operator* $\mathtt{AGG}$ *is monotone and associative, is* GLB-CQA($g()$) *expressible in* AGGR[FOL]*?*

Moreover, if the answer is "yes," it is possible to effectively construct, also in quadratic time, a formula in AGGR[FOL] *that solves* GLB-CQA($g()$).

Open Questions

- Extend our results to aggregate operators that lack monotonicity, associativity, or both.
- Shift from expressibility in AGGR[FOL] to computability in P.

Thanks

Thanks!

References I

Marcelo Arenas, Leopoldo E. Bertossi, and Jan Chomicki.
Scalar aggregation in fd-inconsistent databases.
In *ICDT*, volume 1973 of *Lecture Notes in Computer Science*, pages 39–53. Springer, 2001.

Aziz Amezian El Khalfioui and Jef Wijsen.
Computing range consistent answers to aggregation queries via rewriting.
Proc. ACM Manag. Data, 2(5), November 2024.