

# Extending Codd's Theorem to Databases over Semirings

Carles Noguera

Joint work with Guillermo Badia (Queensland) and Phokion Kolaitis (IBM Almaden)

Prague, February 2026

# Relational Algebra

Edgar (Ted) Codd in 1972 introduced two different query languages for the relational data model.

The first is **Relational Algebra**, which is a procedural language. It is an **algebraic formalism** in which queries are expressed by applying a **sequence of operations** to relations.

A **relational algebra expression** is a string obtained from relation schemas using union, difference, cartesian product, projection, and selection:

$$E := R, S, \dots \mid E_1 \cup E_2 \mid E_1 \setminus E_2 \mid E_1 \times E_2 \mid \pi_L(E) \mid \sigma_{\vartheta}(E)$$

where  $R, S, \dots$  are relations from a schema  $\tau$ ,  $L$  is a list of positions,  $\vartheta$  is a condition.

# Relational Calculus

The second query language is **Relational Calculus**, which is a declarative language. It is a **logical formalism** in which queries are expressed as formulas of **first-order logic**.

A relational calculus expression is an expression of the form  $\{(x_1, \dots, x_k) \mid \varphi(x_1, \dots, x_k)\}$ , where  $\varphi(x_1, \dots, x_k)$  is a first-order formula with  $x_1, \dots, x_k$  as its free variables.

E.g. binary relations  $R(A, B), S(C, D)$ , algebra expression  $\pi_{1,4}(\sigma_{R.B=S.C}(R \times S))$  translates to  $\exists y, z((y = z) \wedge R(x, y) \wedge S(z, w))$ .

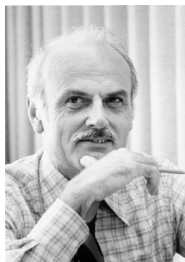
# Relational Calculus under the active domain interpretation

The **active domain** ‘ $\text{adom}(I)$ ’ of a relational database instance  $I$  is the set of all elements that occur in the relations of  $I$ .

Let  $\varphi(x_1, \dots, x_k)$  be a relational calculus formula and let  $I$  be a relational database instance. Then  $\varphi^{\text{adom}(I)}$  is the result of evaluating  $\varphi(x_1, \dots, x_k)$  over  $\text{adom}(I)$  and  $I$ , i.e.,

- all variables and quantifiers are assumed to range over  $\text{adom}(I)$ ;
- the relation symbols in  $\varphi$  are interpreted by the relations in  $I$ .

# Codd's theorem



## Theorem 1

Let  $\tau = (\hat{R}_1, \dots, \hat{R}_p)$  be a schema, and let  $q$  be an  $n$ -ary query. The following statements are equivalent:

1. There is a relational algebra expression  $E$  of arity  $n$  such that  $q^I = E^I$ , for every database  $I$ .
2. There is a relational calculus formula  $\varphi(y_1, \dots, y_n)$  such that  $q^I = \varphi^{\text{adom}(I)}$ , for every database  $I$ .

# Special Classes of Semirings

Let  $\mathbb{K} = (K, +, \cdot, 0, 1)$  be a semiring. We will consider several additional properties.

- $\mathbb{K}$  is *commutative* if the monoid  $(K, \cdot, 1)$  is commutative.
- $\mathbb{K}$  is *zero-sum-free* if for all elements  $a, b \in K$  with  $a + b = 0$ , we have that  $a = 0$  and  $b = 0$ .
- $\mathbb{K}$  has *no zero divisors* if for all elements  $a, b \in K$  with  $a \cdot b = 0$ , we have that  $a = 0$  or  $b = 0$ .
- $\mathbb{K}$  is a *positive* semiring if it is commutative, zero-sum-free, and has no zero divisors.

# Semiring-annotated relations

Assume that  $D$  is a fixed denumerable set, which will be the **domain** of elements occurring in databases.

## Definition 1

Let  $\mathbb{K} = (K, +, \cdot, 0, 1)$  be a semiring and let  $n \geq 0$  be a non-negative integer.

- An  $n$ -ary  $\mathbb{K}$ -relation  $R$  is a function  $R: D^n \rightarrow K$  of finite support, that is to say, the set  $\text{supp}(R) := \{(d_1, \dots, d_n) \in D^n : R(d_1, \dots, d_n) \neq 0\}$  is finite. We call  $n$  the *arity* of  $R$ .
- A  $\mathbb{K}$ -*relation* is an  $n$ -ary  $\mathbb{K}$ -relation  $R$ , for some  $n \geq 0$ .
- If  $n \geq 1$  and  $A \subseteq D$ , then an  $n$ -ary  $\mathbb{K}$ -*relation on*  $A$  is an  $n$ -ary  $\mathbb{K}$ -relation  $R$  with  $\text{supp}(R) \subseteq A^n$ .

# Codd's theorem for $\mathbb{K}$ -databases

**Central Question:** Does Codd's theorem hold for  $\mathbb{K}$ -databases?

**Immediate Obstacles:** What do you do with difference? What do you do with negation?

## Definition 1 (Semiring with monus)

A semiring with monus is a structure  $\mathcal{S} = \langle S, +, \cdot, -, \mathbb{0}, \mathbb{1} \rangle$  where  $\mathcal{S} = \langle S, +, \cdot, \mathbb{0}, \mathbb{1} \rangle$  is a commutative semiring such that:

- (1) The relation  $a \leq b$  defined as  $\exists c(a + c = b)$  is a partial order (known as the **natural order**),
- (2) For all  $a, b$ , there is a smallest  $c$  (denoted  $a - b$ ) such that  $a \leq b + c$ .

# Examples of semirings with monus

## Example 1

- $\mathbb{B} = \langle \{\top, \perp\}, \wedge, \vee, \perp, \top \rangle$ : expand with  $a - b := a \wedge b^*$ .
- $\mathbb{N} = \langle \mathbb{N}, +, \cdot, 0, 1 \rangle$ : expand with  $n \div m$ , defined as  $m - n$  if  $m > n$  and 0 otherwise.
- $\mathbb{L} = \langle L, \wedge, \vee, \perp, \top \rangle$  (complete bounded distributive lattices): expand with ‘pseudo-difference’  $x - y = \min\{z \in L \mid x \vee z \geq y\}$ .
- $\mathbb{F} = \langle [0, 1], \max, \min, 0, 1 \rangle$ : expand with  $a - b = 0$  if  $a \leq b$  and  $a - b = a$  if  $a > b$ .
- $\langle \mathbb{R} \cup \{+\infty\}, \min, +, +\infty, 0 \rangle$ : expand with  $a - b = a$  if  $a <_{\mathbb{R}} b$  and  $a - b = +\infty$  if  $a \geq_{\mathbb{R}} b$ .

# The choice of difference

## Why semirings with monus?

- They comprise a large class of semirings.
- Geerts & Poggi 2009: relational algebra with difference (**m-semirings**). The failure of some useful identities on some semirings with monus is criticized by Amsterdamer, Deutch & Tannen 2011.
- Bag semantics is the default SQL semantics. Monus is also used in bag semantics. Identities like  $R \cup R = R$  fail on the bag semiring.

# The choice of negation

We do not want to limit the programmer to use formulas in negation normal form.

To get around this, we opt to use the binary connective  $\rightarrow$  ('but not'), a limited form of negation. Long history in the literature on non-classical logics:

- Cecylia Rauszer. Semi-Boolean algebras and their applications to intuitionistic logic with dual operations. *Fundamenta Mathematicae*, 83(3):219–249, 1974.
- Nicolas D. Goodman. The logic of contradiction. *Mathematical Logic Quarterly*, 27(8–10):119–126, 1981.

Criticized in [Grädel & Tannen 2024](#) when used to define an unrestricted negation  $1 \rightarrow \varphi$ .

# Basic Relational Algebra ( $\mathcal{BRA}$ ) over semirings

Introduce the operations **Union**, **Difference**, **Cartesian Product**, **Projection**, and **Selection** on  $\mathbb{K}$ -relations e.g.

- **Difference:** If  $R_1$  and  $R_2$  are two  $n$ -ary  $\mathbb{K}$ -relations, then the *difference*  $R_1 \setminus R_2$  is the  $n$ -ary  $\mathbb{K}$ -relation defined by  $(R_1 \setminus R_2)(\mathbf{t}) = R_1(\mathbf{t}) - R_2(\mathbf{t})$ , for all  $\mathbf{t} \in D^n$ .

Must add a **Support** operation.

- **Support:** If  $R$  is an  $n$ -ary  $\mathbb{K}$ -relation, then the *support*  $\text{supp}(R)$  of  $R$  is the  $n$ -ary  $\mathbb{K}$ -relation defined by  $\text{supp}(R)(\mathbf{t}) = \mathbf{s}(R(\mathbf{t}))$ , for all  $\mathbf{t} \in D^n$ .

## Definition 2 (Support semiring)

$\mathcal{S} = (S, +, \cdot, \mathbb{0}, \mathbb{1}, \mathbf{s})$  is a **support semiring** if  $\mathcal{S} = (S, +, \cdot, \mathbb{0}, \mathbb{1})$  is a semiring and, for each  $a \in S$ , we have

$$\mathbf{s}(a) = \begin{cases} 1 & \text{if } a \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

# Semiring-annotated structures

## Definition 3 ( $\mathbb{K}$ -structures and assignments)

Let  $\mathbb{K} = (K, +, \cdot, -, \mathbf{s}, 0, 1)$  be a semiring with monus and support, and let  $\tau = (\hat{R}_1, \dots, \hat{R}_p)$  be a schema.

- A  $\mathbb{K}$ -structure over  $\tau$  is a tuple  $\mathbf{A} = (A, R_1, \dots, R_p)$ , where  $A$  is a non-empty subset of  $D$ , called the *universe* of  $\mathbf{A}$ , and each  $R_i$  is a  $\mathbb{K}$ -relation on  $A$  whose arity matches the arity of the relation symbol  $\hat{R}_i$ ,  $1 \leq i \leq p$ .

We say that  $\mathbf{A}$  is a *finite*  $\mathbb{K}$ -structure if its universe  $A$  is a finite set.

- An *assignment* on  $\mathbf{A}$  is a function  $\alpha: V \rightarrow A$ , assigning to every variable an element in  $A$ .
- If  $\alpha$  is an assignment on  $\mathbf{A}$  and  $b$  is an element of  $A$ , then we write  $\alpha[x_i/b]$  for the assignment  $\beta$  on  $\mathbf{A}$  such that  $\beta(x_i) = b$  and  $\beta(x_j) = \alpha(x_j)$ , for every variable  $x_j \neq x_i$ .

# Semantics of Basic Relational Calculus ( $\mathcal{BRC}$ )

## Definition 4

The **semantics of  $\varphi$  on  $\mathbb{K}$**  is a function  $\|\varphi\|$  that takes as input a  $\mathbb{K}$ -structure  $\mathbf{A} = (A, R_1, \dots, R_p)$  and an assignment  $\alpha$  on  $\mathbf{A}$ , and returns a value  $\|\varphi\|(\mathbf{A}, \alpha)$  in  $\mathbb{K}$  that is defined recursively and simultaneously for all assignments as follows:

- $\|x_i = x_j\|(\mathbf{A}, \alpha) = \begin{cases} 1 & \text{if } \alpha(x_i) = \alpha(x_j); \\ 0 & \text{otherwise.} \end{cases}$
- $\|\hat{R}_i(x_{i_1}, \dots, x_{i_n})\|(\mathbf{A}, \alpha) = R_i(\alpha(x_{i_1}), \dots, \alpha(x_{i_n}))$ , for  $1 \leq i \leq p$ .
- $\|\varphi_1 \bar{\circ} \varphi_2\|(\mathbf{A}, \alpha) = \|\varphi_1\|(\mathbf{A}, \alpha) \circ \|\varphi_2\|(\mathbf{A}, \alpha)$  where  $\bar{\circ} \in \{\wedge, \vee, \neg\}$  and  $\circ$  is the corresponding operation from  $\{\cdot, +, -\}$ .
- $\|\nabla \varphi\|(\mathbf{A}, \alpha) = s(\|\varphi\|(\mathbf{A}, \alpha))$ .
- $\|\exists x_i \varphi\|(\mathbf{A}, \alpha) = \sum_{b \in A} \|\varphi\|(\mathbf{A}, \alpha[x_i/b])$ .

## Definition 5

Let  $\mathbb{K}$  be a semiring with monus and support, and let  $\tau$  be a schema.

- If  $\varphi(y_1, \dots, y_n)$  is a  $\mathcal{BRC}(\tau)$ -formula whose free variables are  $y_1, \dots, y_n$  and if  $\mathbf{A}$  is a  $\mathbb{K}$ -structure, then  $\varphi^{\mathbf{A}}: D^n \rightarrow K$  is the  $n$ -ary  $\mathbb{K}$ -relation defined as follows:

$$\varphi^{\mathbf{A}}(a_1, \dots, a_n) = \begin{cases} \|\varphi\|(\mathbf{A}, (a_1, \dots, a_n)) & \text{if } (a_1, \dots, a_n) \in A^n \\ 0 & \text{if } (a_1, \dots, a_n) \notin A^n \end{cases}$$

## Definition 6

- The **active domain** of  $\mathbf{I} = (R_1, \dots, R_p)$ , denoted by  $\text{adom}(\mathbf{I})$  contains all elements of  $D$  that occur in the support of at least one relation  $R_i$ .
- $\mathbf{A}(\mathbf{I})$  is the  $\mathbb{K}$ -structure with universe the active domain of  $\mathbf{I}$  and with relations those of  $\mathbf{I}$ . In symbols, we have  $\mathbf{A}(\mathbf{I}) = (\text{adom}(\mathbf{I}), R_1, \dots, R_p)$ .

## Theorem 2

Let  $\mathbb{K}$  be a semiring with monus and support, let  $\tau = (\hat{R}_1, \dots, \hat{R}_p)$  be a schema, and let  $q$  be an  $n$ -ary query. The following statements are equivalent:

1. There is a  $\mathcal{BRA}(\tau)$ -expression  $E$  of arity  $n$  such that  $q^{\mathbf{I}} = E^{\mathbf{I}}$ , for every  $\mathbb{K}$ -database  $\mathbf{I}$ .
2. There is a  $\mathcal{BRC}(\tau)$ -formula  $\varphi(y_1, \dots, y_n)$  such that  $q^{\mathbf{I}} = \varphi^{\mathbf{A}(\mathbf{I})}$ , for every  $\mathbb{K}$ -database  $\mathbf{I}$ .

## Adding division to basic relation algebra

( $R = [\text{STUDENT}, \text{COURSE}]$ ,  $S = [\text{BAHAREH}' \text{ COURSES}]$ ); retrieve the set of students that attend **all** of Bahareh' courses:  $R \div S$ .

### Definition 7 (Division)

Let  $R_1 : D^{n_1} \rightarrow K$  and  $R_2 : D^{n_2} \rightarrow K$  be two  $\mathbb{K}$ -relations with  $n_1 > n_2$ . The *division* of  $R_1$  and  $R_2$  is the  $\mathbb{K}$ -relation  $R_1 \div R_2 : D^{n_1 - n_2} \rightarrow K$  defined by

$$R_1 \div R_2(\mathbf{a}) = s\left(\sum_{\mathbf{b} \in D^{n_2}} R_1(\mathbf{a}, \mathbf{b})\right) \cdot \prod_{\mathbf{b} : R_2(\mathbf{b}) \neq 0} R_1(\mathbf{a}, \mathbf{b}).$$

We stipulate that the empty product is equal to 1. Observe that  $R_1 \div R_2 = R_1 \div \text{supp}(R_2)$ .

# Adding division to basic relational algebra

## Example 2

If  $\mathbb{B} = (\{0, 1\}, \wedge, \vee, 0, 1)$  is the Boolean semiring, then the division operation introduced here coincides with the division (also known as *quotient*) operation of relational algebra, i.e., if  $R_1$  and  $R_2$  are ordinary relations with  $\text{arity}(R_2) < \text{arity}(R_1)$ , then  $R_1 \div R_2(\mathbf{a}) = 1$  if and only if for all  $\mathbf{b}$  with  $R_2(\mathbf{b}) = 1$ , we have that  $R_1(\mathbf{a}, \mathbf{b}) = 1$ .

## Example 3

Let  $\mathbb{K} = (K, +_{\mathbb{K}}, \cdot_{\mathbb{K}}, 0_{\mathbb{K}}, 1_{\mathbb{K}})$  be a positive semiring with monus and support. Assume that  $R_1$  and  $R_2$  are two  $\mathbb{K}$ -relations with  $\text{arity}(R_2) < \text{arity}(R_1)$ , and let  $\mathbf{a}$  be a tuple in  $D^{n_1 - n_2}$ .

- 1 If  $\mathbf{a} \in \text{supp}(\pi_{1\dots n_1 - n_2}(R_1))$  and for every tuple  $\mathbf{b} \in \text{supp}(R_2)$  we have that  $(\mathbf{a}, \mathbf{b}) \in \text{supp}(R_1)$ , then the following hold:
  - 1 If  $\mathbb{K}$  is the bag semiring  $(\mathbb{N}, +, \cdot, 0, 1)$ , then  $R_1 \div R_2(\mathbf{a}) = \prod_{R_2(\mathbf{b}) \neq 0} R_1(\mathbf{a}, \mathbf{b})$ .
  - 2 If  $\mathbb{K}$  is the tropical semiring  $(\mathbb{N} \cup \{\infty\}, \min, +, \infty, 0)$  of the natural numbers, then  $R_1 \div R_2(\mathbf{a}) = \sum_{R_2(\mathbf{b}) \neq 0} R_1(\mathbf{a}, \mathbf{b})$ .
  - 3 If  $\mathbb{K}$  is the fuzzy semiring  $([0, 1], \max, \min, 0, 1)$ , then  $R_1 \div R_2(\mathbf{a}) = \min_{R_2(\mathbf{b}) \neq 0} R_1(\mathbf{a}, \mathbf{b})$ .

# Independence of division

## Theorem 3

Let  $\mathbb{N} = (N, +, \cdot, \div, s, 0, 1)$  be the bag semiring expanded with the truncated subtraction  $\div$  as monus and with support  $s$ . Then **there is no expression of *BR* that defines the division operation  $\div$  on  $\mathbb{N}$ -databases** (i.e., division is not expressible using union, difference, Cartesian product, projection, selection, and support on  $\mathbb{N}$ -databases).

## Hint of proof

Let  $\tau = (R, S)$  be a schema consisting of a binary relation symbol  $R$  and a unary relation symbol  $S$ . For every  $n \geq 1$ , let  $a, b_1, \dots, b_n$  be distinct elements from the domain  $D$  and let  $\mathbf{I}(n) = (R_n, S_n)$  be the bag database ( $\mathbb{N}$ -database), where  $R_n$  and  $S_n$  are the following bags ( $\mathbb{N}$ -relations):

- $R_n(a, b_i) = 2$ , for  $1 \leq i \leq n$ , while  $R_n(c, d) = 0$ , for all other pairs of elements of  $D$ .
- $S_n(b_i) = 1$ , for  $1 \leq i \leq n$ , while  $S_n(c) = 0$ , for every other element of  $D$ .

Clearly, for every  $n \geq 1$ , the division  $R_n \div S_n$  is the bag such that  $(R_n \div S_n)(a) = 2^n$ , while  $(R_n \div S_n)(c) = 0$ , for all other elements of  $D$ .

One can show that no  $\mathcal{BR}\mathcal{A}$ -expression defines  $R \div S$  by showing that, for sufficiently large  $n$ , no such expression involving  $R_n$  and  $S_n$  can have an element of multiplicity  $2^n$ .

# Relational Algebra and Relational Calculus

$$\mathcal{RA} = \mathcal{BRA} + \text{division}$$

Add:

- If  $E_1, E_2$  are  $\mathcal{RA}(\tau)$ -expressions of arities  $n_1, n_2$ , respectively, such that  $n_1 > n_2$ , then  $E_1 \div E_2$  is an  $\mathcal{RA}(\tau)$ -expression of arity  $n_1 - n_2$ .

$$\mathcal{RC} = \mathcal{BRC} + \text{universal quantifier}$$

Add the following clause to the semantics:

- $\|\forall x_i \varphi\|(\mathbf{A}, \alpha) = \prod_{b \in A} \|\varphi\|(\mathbf{A}, \alpha[x_i/b])$ .

## Second Codd theorem

### Theorem 4

Let  $\mathbb{K}$  be a positive semiring with monus and support, let  $\tau = (\hat{R}_1, \dots, \hat{R}_p)$  be a schema, and let  $q$  be an  $n$ -ary query. The following statements are equivalent:

1. There is an  $\mathcal{RA}(\tau)$ -expression  $E$  of arity  $n$  such that  $q^{\mathbf{I}} = E^{\mathbf{I}}$ , for every  $\mathbb{K}$ -database  $\mathbf{I}$ .
2. There is an  $\mathcal{RC}(\tau)$ -formula  $\varphi(y_1, \dots, y_n)$  such that  $q^{\mathbf{I}} = \varphi^{\mathbf{A}(\mathbf{I})}$ , for every  $\mathbb{K}$ -database  $\mathbf{I}$ .

## Hint of proof

(1)  $\implies$  (2) : (induction)

- E.g.  $\varphi$  is of the form  $\forall y_n \psi$
- take  $E_\varphi$  to be the  $\mathcal{RA}$ -expression

$$E_\psi \div E_{\text{adom}}.$$

(2)  $\implies$  (1) :

- $E$  is of the form  $E_1 \div E_2$ ,
- where  $\eta$  is  $\nabla \exists y_1 (y_1 = y_1)$ , let

$$(\nabla \exists y_{n_1-n_2+1} \dots \exists y_{n_1} \varphi_1) \wedge (\forall y_{n_1-n_2+1} \dots \forall y_{n_1} ((\eta \leftarrow \nabla \varphi_2) \vee (\nabla \varphi_2 \wedge \varphi_1))).$$

# Domain independence

## Definition 2

Given a  $\mathbb{K}$ -database  $\mathbf{I} = (R_1, \dots, R_p)$ , we write  $\mathbf{A}(\mathbf{I})$  to denote the  $\mathbb{K}$ -structure with universe the active domain of  $\mathbf{I}$  and with relations those of  $\mathbf{I}$ . In symbols, we have  $\mathbf{A}(\mathbf{I}) = (\text{adom}(\mathbf{I}), R_1, \dots, R_p)$ .

## Definition 3 (Domain Independence)

A  $\mathcal{RC}$ -formula  $\varphi$  is *domain independent* if for every  $\mathbb{K}$ -database  $\mathbf{I} = (R_1, \dots, R_p)$  and every  $\mathbb{K}$ -structure  $\mathbf{B} = (B, R_1, \dots, R_p)$  with  $\text{adom}(\mathbf{I}) \subseteq B \subseteq D$ , we have  $\varphi^{\mathbf{B}} = \varphi^{\mathbf{A}(\mathbf{I})}$ .

# Codd's theorem with domain independence

## Theorem 5

Let  $\mathbb{K}$  be a positive semiring with monus and support, let  $\tau = (\hat{R}_1, \dots, \hat{R}_p)$  be a schema, and let  $q$  be an  $n$ -ary query. The following statements are equivalent:

1. There is an  $\mathcal{RA}(\tau)$ -expression  $E$  of arity  $n$  such that  $q^{\mathbf{I}} = E^{\mathbf{I}}$ , for every  $\mathbb{K}$ -database  $\mathbf{I}$ .
2. There is a domain independent  $\mathcal{RC}(\tau)$ -formula  $\varphi(y_1, \dots, y_n)$  such that  $q^{\mathbf{I}} = \varphi^{\mathbf{A}(\mathbf{I})}$ , for every  $\mathbb{K}$ -database  $\mathbf{I}$ .
3. There is an  $\mathcal{RC}(\tau)$ -formula  $\varphi(y_1, \dots, y_n)$  such that  $q^{\mathbf{I}} = \varphi^{\mathbf{A}(\mathbf{I})}$ , for every  $\mathbb{K}$ -database  $\mathbf{I}$ .

# Open problems

Two versions of Codd's Theorem over semirings (one adding division and the universal quantifier).

The five basic operations of relational algebra cannot express the division operation on the bag semiring.

## Open problems:

- In which semirings can  $\mathcal{BRA}$  express division?
- [Suciu 2024](#) 'Different Differences in Semirings'. Are there versions of Codd's Theorem when alternative choices for the difference operation and for negation are made?