Relational Algebra

Selection and Projection

Yihao Sun Utah State University Sept. 2026

Theater Addresses

Which theaters show a film, and what are their addresses?

M (Movies)

Title	Director	Actor
filmA	Bergman	actor1
filmA	Bergman	actor2
filmB	Other	actor3

P (Pariscope)

Theater	Title	Schedule
Champo	filmA	18:00
Champo	filmA	21:00
Rialto	filmB	20:00

L (Location)

Theater	Address	Phone
Champo	addr1	phone1
Rialto	addr2	phone2

What is the CQ for this question?

A Sequential Join

$$\text{ans}(h, a) \leftarrow M(t, d, c),\ P(h, t, s),\ L(h, a, p)$$

$$T = M \bowtie P$$

Title	Director	Actor	Theater	Schedule
filmA	Bergman	actor1	Champo	18:00
filmA	Bergman	actor1	Champo	21:00
filmA	Bergman	actor2	Champo	18:00
filmA	Bergman	actor2	Champo	21:00
filmB	Other	actor3	Rialto	20:00

5 tuples. Four differ only in actor or showtime.

Complete the Join

$J(t, d, c, h, s, a, p) \leftarrow T(t, d, c, h, s),\ L(h, a, p)$

$J = T \bowtie L = (M \bowtie P) \bowtie L$

5 tuples, 7 columns.

Title	Director	Actor	Theater	Schedule	Address	Phone
filmA	Bergman	actor1	Champo	18:00	addr1	phone1
filmA	Bergman	actor1	Champo	21:00	addr1	phone1
filmA	Bergman	actor2	Champo	18:00	addr1	phone1
filmA	Bergman	actor2	Champo	21:00	addr1	phone1
filmB	Other	actor3	Rialto	20:00	addr2	phone2

$\text{ans}(h, a) \leftarrow J(t, d, c, h, s, a, p)$ (requested pairs)

One Relation, Two Roles

$$\text{ans}(x, z) \leftarrow \text{edge}(x, y),\ \text{edge}(y, z)$$

$E = I(\text{edge})$ two directed edges

src	dst
Logan	Ogden
Ogden	SLC

$$E \bowtie E = E$$

Natural join checks every shared column name.

How do we make only the middle variable shared?

Renaming Makes the Roles Explicit

$$E_1 = \delta_{\text{src dst} \to x\, y}(E)$$

x	y
Logan	Ogden
Ogden	SLC

$$E_2 = \delta_{\text{src dst} \to y\, z}(E)$$

y	z
Logan	Ogden
Ogden	SLC

$$E_1 \bowtie E_2 = \{\langle x : \text{Logan}, y : \text{Ogden}, z : \text{SLC} \rangle\}$$

Rename columns. Preserve values. Share only y.

What Does the Next Join Need?

$$\pi_K(T) = \{\, v[K] \mid v \in T \,\}$$

$$\pi_{\text{Theater}}(T) = \{\text{Champo}, \text{Rialto}\}$$

Keep Theater. Forget Title, Director, Actor, and Schedule.

Different witnesses can become the same retained valuation.

Fewer Witnesses, Less Work

$T \bowtie L$: 5 tuples, 7 columns

$\pi_{\text{Theater}}(T) \bowtie L$: 2 tuples, 3 columns

The same theater does not need a separate address lookup for every witness.

$$\pi_{\text{Theater,Address}}(T \bowtie L) = \pi_{\text{Theater,Address}}(\pi_{\text{Theater}}(T) \bowtie L)$$

Which Variables Must Survive?

Keep variables needed by the answer, remaining atoms, or pending conditions.

$$U = \mathrm{sort}(E), \qquad V = \mathrm{sort}(F)$$

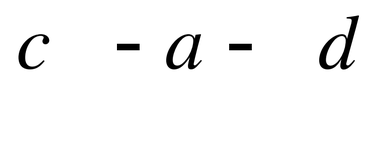
$$\frac{(U \cap V) \cup (H \cap U) \subseteq K \subseteq U \qquad H \subseteq U \cup V}{\pi_H(E \bowtie F) \equiv \pi_H(\pi_K(E) \bowtie F)}$$

Read the condition first. Then decide whether the rewrite is legal.

CQ Zoo: Star

$$\text{ans}(b, d) \leftarrow R(a, b),\ S(a, c),\ T(a, d)$$

$$R(a, b) = \{(1,10)\}$$
$$S(a, c) = \{(1,20), (1,21)\}$$
$$T(a, d) = \{(1,30), (1,31)\}$$



$$\underbrace{R \bowtie S}_{\text{already computed}} \bowtie T$$

Which columns must survive? How many rows enter and leave the next join?

CQ Zoo: Cycle

$$\text{ans}(d) \leftarrow R(a,b),\ S(b,c),\ T(c,d),\ U(d,a)$$

$$R(a,b) = \{(1,2)\}$$
$$S(b,c) = \{(2,3)\}$$
$$T(c,d) = \{(3,4),(3,5)\}$$
$$U(d,a) = \{(4,1),(5,9)\}$$

$$\begin{array}{ccc} a & - & b \\ | & & | \\ d & - & c \end{array}$$

$J = (R \bowtie S) \bowtie T$ Only U remains.

The answer returns only the endpoint. Can we forget the starting variable?

A Bound Title Selects Showings

$u \in M$ one chosen tuple

Title	Director	Actor
filmA	Bergman	actor1

Which showings agree with its Title?

$$\sigma_{\text{Title}=u(\text{Title})}(P) = \{v \in P \mid v(\text{Title}) = u(\text{Title})\} = \sigma_{\text{Title}=\text{filmA}}(P)$$

Theater	Title	Schedule
Champo	filmA	18:00
Champo	filmA	21:00

The Valuation Supplies the Next Test

$$u \bowtie v, \qquad v(\text{Theater}) = \text{Champo}$$

$$\sigma_{\text{Theater}=v(\text{Theater})}(L) = \sigma_{\text{Theater}=\text{Champo}}(L)$$

Theater	Address	Phone
Champo	addr1	phone1

$$J = M \bowtie P \bowtie L$$
$$= \{\, u \bowtie v \bowtie w \mid u \in M,\ v \in \sigma_{\text{Title}=u(\text{Title})}(P),\ w \in \sigma_{\text{Theater}=v(\text{Theater})}(L) \,\}$$

All compatible choices give the same full join.

Selection Tests a Valuation

Now the query itself requires a Bergman film.

$$\sigma_{\varphi}(E) = \{\, v \in E \mid v \vDash \varphi \,\}$$

$$\sigma_{\text{Director=Bergman}}(M)$$

Title	Director	Actor
filmA	Bergman	actor1
filmA	Bergman	actor2

Fewer rows. The same columns.

Test the Condition Earlier

$\varphi \ : \ \text{Director} = \text{Bergman}$

$$\sigma_{\varphi}(M \bowtie P) \equiv \sigma_{\varphi}(M) \bowtie P$$

$$\frac{\text{vars}(\varphi) \subseteq \text{sort}(E)}{\sigma_{\varphi}(E \bowtie F) \equiv \sigma_{\varphi}(E) \bowtie F}$$

The other film never enters the join.

Forget Director only after its condition has been checked.

SPJ Plan for Our First Instance

$$B = \pi_{\text{Title}}(\sigma_{\text{Director=Bergman}}(M))$$

$$C = \pi_{\text{Theater}}(B \bowtie P)$$

$$Q = \pi_{\text{Theater,Address}}(C \bowtie L)$$

Each intermediate relation contains only what the remaining query needs.

Theater	Address
Champo	addr1

Reading an Algebra Expression

$$E ::= R \mid \sigma_{\varphi}(E) \mid \pi_{K}(E) \mid E \bowtie E \mid \delta_{f}(E)$$

$$\frac{E : U \qquad K \subseteq U}{\pi_{K}(E) : K} \qquad \frac{E : U \qquad \mathrm{vars}(\varphi) \subseteq U}{\sigma_{\varphi}(E) : U}$$

$$\frac{E : U \qquad F : V}{E \bowtie F : U \cup V} \qquad \frac{E : U \quad f \text{ one-to-one on } U}{\delta_{f}(E) : f(U)}$$

SPJR: selection, projection, join, and renaming.

CQ Zoo: Path + a Condition

$$\text{ans}(a, d) \leftarrow R(a, b),\ S(b, c),\ T(c, d),\ b < c$$

$$a \text{ - } b \text{ - } c \text{ - } d$$

$$J = R \bowtie S$$

$\pi_{a,c}(J)$ before testing $b < c$?

Write a correct SPJ plan. Mark the last use of every variable.

One Query, Several Correct Plans

$H = \{\text{Theater}, \text{Address}\}, \qquad \varphi : \text{Director} = \text{Bergman}$

$$\pi_H\Big(\sigma_\varphi((M \bowtie P) \bowtie L)\Big)$$

$$\pi_H\Big(\pi_{\text{Theater}}(\pi_{\text{Title}}(\sigma_\varphi(M)) \bowtie P) \bowtie L\Big)$$

Equivalent answers. Different intermediate relations.

The algebra gives us rules for building and changing a computation.

Same Column, Different Frequencies

New illustrative instance: 100 Movies tuples.

Director	Tuples
Director A	80
Bergman	2
Director B	18

$$\text{sel}_{A=c}(R) = \frac{|\sigma_{A=c}(R)|}{|R|}$$

$$\text{sel}_{\text{Director=Bergman}}(M) = \frac{2}{100}$$

Skew means unequal value frequencies.

What Does Early Selection Save?

One Movies tuple per title; exactly three showings per title.

$$\sigma_{\text{Director=Bergman}}(M \bowtie P) \equiv \sigma_{\text{Director=Bergman}}(M) \bowtie P$$

Intermediate	Join, then select	Select, then join
Movies tuples fed to join	100	2
Join tuples produced	300	6
Qualifying join tuples	6	6

But finding those two Movies tuples may still require reading all 100.

An Index Remembers Where to Look

An extra structure maps a column value to matching row locations.

$$\sigma_{\text{Director=Bergman}}(M) = \{\, v \in M \mid v(\text{Director}) = \text{Bergman} \,\}$$

Index on Director

Director value	Row IDs
Bergman	17, 83



Retrieved Movies tuples

Row ID	Title	Director	Actor
17	film17	Bergman	actor17
83	film83	Bergman	actor83

Row IDs are storage metadata.

The selection result is unchanged. The way we find it has changed.

An Index Still Returns Every Match

Assume the index already exists.

Director value	Scan: tuples tested	Index: matching row IDs
Bergman	100	2
Director A	100	80

A frequent value still returns many matches; a scan may be cheaper.

t is bound: $\sigma_{\text{Title}=\nu(t)}(P)$

Which column should the next lookup use?