

# Incremental Maintenance for Leapfrog Triejoin

Todd Veldhuizen\*

## Abstract

We present an incremental maintenance algorithm for leapfrog triejoin. The algorithm maintains rules in time proportional (modulo log factors) to the edit distance between leapfrog triejoin traces.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Aspiration . . . . .	3
1.2	Summary of the maintenance algorithm . . . . .	4
1.3	Background, terminology, and notations . . . . .	4
<b>2</b>	<b>Maintaining head predicates</b>	<b>5</b>
2.1	Projection-free rules . . . . .	5
2.2	Rules with projection . . . . .	6
2.3	Aggregations . . . . .	6
<b>3</b>	<b>Algorithms &amp; Data structures</b>	<b>8</b>
3.1	Scans . . . . .	8
3.2	Interval trees . . . . .	12
3.3	Delta-iterators . . . . .	13
<b>4</b>	<b>Maintaining rule bodies</b>	<b>13</b>
4.1	Sensitivity indices . . . . .	14
4.2	Sensitivity indices for predicates with multiple arguments . . . . .	16

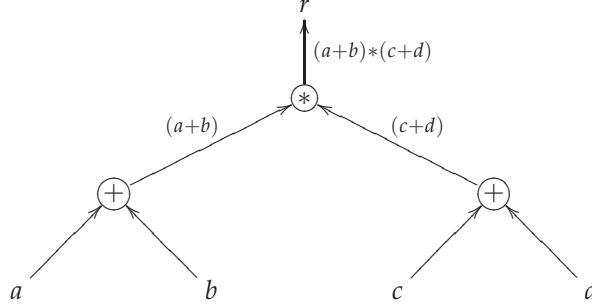
---

\*LogicBlox Inc., tveldhui@acm.org

# 1 Introduction

Incremental evaluation is a perennial topic in computer science. The basic problem is easily described: given an expensive computation, and some change to its inputs, we want to efficiently update the result, without recomputing from scratch.

The most common traditional approach to such problems is to consider a computation graph, in which edges carry values, and vertices represent small computations, inputs or outputs. For example, the arithmetic expression  $r = (a + b) * (c + d)$  could be represented by this graph:



We can regard the above graph as a *trace*, i.e., a low-level history of the computation. When an input such as  $b$  changed, the effect of the change can be rippled through the graph, only re-evaluating those vertices whose inputs change. We can call this process *trace maintenance*, and if done properly, can be done in time cost proportional to the number of changes required to update the trace.

In databases the problem of incremental evaluation is known as *incremental view maintenance*. A view is simply a query installed in a database, and kept up-to-date as its input relations change.

Incremental view maintenance has historically been done using one of two techniques [4]:

1. Syntactic approaches derive special rules to update a view. For example, given a rule such as:

$$C(x) \leftarrow A(x), B(x).$$

which computes the intersection  $A \cap B$ , one can automatically derive rules to update the view when elements are inserted to  $A$  or  $B$ . For example, a rule which says ‘If  $x$  is inserted to  $A$ , and  $x \in B$ , then insert  $x$  into  $C$ ’ can be written:

$$+C(x) \leftarrow +A(x), B(x)$$

There are two challenges associated with this approach: (a) for complex rules one can encounter a combinatorial explosion of update rules; and (b) the update rules may be difficult to evaluate efficiently. In particular, any claim of efficiency for this approach must resort to a *deus ex machina* appeal to the strength of the query optimizer.

2. Algebraic approaches follow a trace-maintenance approach, at a coarse level of granularity, where each vertex in the graph represents an algebra operator (e.g., join, projection). One defines special maintenance algorithms for each operator, so that e.g. a projection can be maintained efficiently when its input relation is updated.

In this paper we present an incremental maintenance algorithm for leapfrog triejoin [6], a join algorithm with worst-case optimality guarantees. This maintenance algorithm is implemented in the Delve runtime engine of our commercial Datalog system LogicBlox<sup>®</sup>.

Our approach to incremental maintenance is rather different from the usual database approaches, and is loosely inspired by the dynamization procedure of Acar et al [1]. It hews to the *trace maintenance* approach, arguably the traditional technique in computer science. Unlike the algebraic approaches mentioned above, we maintain the trace at a very fine level of granularity—at the level of individual iterator operations in the leapfrog triejoin algorithm. Our maintenance algorithm has time cost proportional to trace distance (modulo log factors), giving it an optimality guarantee.

## 1.1 Aspiration

Suppose we have some Datalog rule, for example:

$$F(x, y) \leftarrow G(x, z), H(y, z), I(x, y, z).$$

After evaluating this rule to calculate  $F(x, y)$ , some transaction(s) are committed that modify  $G$ ,  $H$ , and/or  $I$ , and we wish to update  $F$  to reflect these changes. In many situations it is prohibitively expensive to recalculate  $F$  from scratch, so we instead aim to efficiently *maintain*  $F(x, y)$  based on the changes made to the predicates.

The aspiration of our maintenance algorithm is *maintenance cost proportional to trace edit distance*. Unpacking this a bit:

- By maintenance cost, we mean the number of steps required to maintain a rule, i.e. produce new versions of the head predicates in response to changes made to body predicates.
- By trace, we mean a low-level step-by-step description of the operations performed during full evaluation of a rule, i.e., a succinct history of the computation. We maintain traces at the level of predicate iterator operations, so these steps might include items such as “position the iterator for  $G(x, z)$  at a least upper bound for  $(x = 1531, z = 142)$ ”.
- By *trace edit distance*, we mean comparing side-by-side the trace for full-evaluation on the original predicates (e.g.  $G, H, I$ ) with the trace for full-evaluation on the modified predicates (e.g.  $G', H', I'$ ), and counting how many changes must be made to the original trace to turn it into the trace for full-evaluation on the modified predicates.

For a trivial illustration, suppose I have a rule

$$C[x]=z \leftarrow z=A[x]+B[x].$$

If I evaluate this rule in a hypothetical debugging mode where each step is logged, I might get a table like this:

x	A[x]	B[x]	C[x]
0	0	0	0
1	10	0	10
2	0	1	1
3	30	1	31
4	0	0	0
5	0	0	0

This table is a rough approximation of what we mean by a ‘trace’: a step-by-step description of the evaluation. Now suppose I make some changes to  $A[x]$  and do another full evaluation, which produces this table (with differences marked by an asterisk):

x	A[x]	B[x]	C[x]
0	0	0	0
1	10	0	10
2	20*	1	21*
3	30	1	31
4	0	0	0
5	50*	0	50*

If I run the unix command ‘diff’ on these two tables, I get:

< 2		0	1	1
> 2		20	1	21
< 5		0	0	0
> 5		50	0	50

The length of this diff hints at what we mean by ‘trace edit distance’: the number of changes (edits) you’d need to make to the original trace to turn it into the trace of full-evaluation on the modified predicates.

There is a large history to the general problem of incremental maintenance (not just for Datalog) that says this goal is achievable; the challenge is finding a solution that achieves this goal yet *performs well*.

## 1.2 Summary of the maintenance algorithm

We give a brief sketch of our approach, for orientation.

First, some notations. Given two versions  $C, C'$  of a predicate  $C(x)$ , we write  $(C \cdots C')$  for the *difference* between the two versions. We mean by this a ‘delta’ relation of the form:

$$(C \cdots C')(x, \Delta) = \begin{cases} (x, \text{INSERT}) & \text{if } x \notin C \text{ and } x \in C' \\ (x, \text{ERASE}) & \text{if } x \in C \text{ and } x \notin C' \end{cases}$$

i.e. a predicate enumerating the differences between  $C, C'$ , with the  $\Delta$  variable taking on values INSERT and ERASE.

Consider the rule mentioned above:

$$F(x, y) \leftarrow G(x, z), H(y, z), I(x, y, z).$$

Let  $\text{Body}[G, H, I](x, y, z) = G(x, z), H(y, z), I(x, y, z)$  be the body of the rule. The basic approach to maintenance is to evaluate a rule of the form:

$$\delta F(x, y, \Delta) \leftarrow (\text{Body}[G, H, I] \cdots \text{Body}[G', H', I'])(x, y, z, \Delta), \text{ChangeOracle}(x, y, z).$$

- The left-hand side  $\delta F(x, y, \Delta)$  gives a set of changes to be applied to the predicate  $F$  to produce the updated predicate  $F'$ .
- The term  $(\text{Body}[G, H, I] \cdots \text{Body}[G', H', I'])(x, y, z, \Delta)$  enumerates differences in the satisfying assignments of the rule-body for  $G, H, I$  vs.  $G', H', I'$ . This is done by simply evaluating both bodies and comparing the results.
- The  $\text{ChangeOracle}(x, y, z)$  term serves to restrict evaluation to just those regions of the tuple-space where changes might occur. The maintenance rule would be correct (but inefficient) if  $\text{ChangeOracle}(x, y, z)$  were omitted.

There are two primary tasks:

1. How to represent head predicates, and how to update them given deltas. We describe these techniques in Section 2, and some specialized data structures and algorithms in Section 3.
2. How to construct and employ the  $\text{ChangeOracle}$  predicate. This is described in Section 4.

## 1.3 Background, terminology, and notations

Our variant of Datalog supports both relations such as  $R(x_1, \dots, x_k)$  and functions such as  $F[x_1, \dots, x_k] = y$ . We refer to functions and relations as *predicates*. A relation or function symbol together with its arguments is an *atom*.

A Datalog rule is written in the form *head*  $\leftarrow$  *body*. A rule head contains one or more atoms; a body is a first-order formula. Users write rules in a relaxed form without quantifiers, for example:

$$S(x, y) \leftarrow A(x, y), B(y, z).$$

Internally, rules are represented in the more restricted form of Figure 1, with explicit quantifiers; for example:

$$\forall x, y. S(x, y) \leftarrow \exists z. A(x, y), B(y, z)$$

```

conj ::= [  $\exists \vec{x} .$  ] dform [ , dform  $\dots$  ]
dform ::= atom | disj | negation
atom ::=  $R(\vec{y})$  |  $F[\vec{y}] = \vec{z}$ 
disj ::= conj; conj[; conj  $\dots$  ]
negation ::= !conj

rule ::=  $\forall \vec{x} .$  head  $\leftarrow$  conj
head ::= atom [ , atom  $\dots$  ]

```

Fig. 1: Internal representation of rules.

Variables occurring in both head and body are placed in the rule-level universal quantifier block (e.g.  $x, y$ ); variables occurring only in the body are ascribed to the existential quantifier block of the smallest conjunction (conj) encompassing their uses (e.g.  $z$ ).

A *materialized predicate* is one whose elements are stored in a data structure. Materialized predicates can be either *extensional* (EBD) or *intensional* (IDB):

- An extensional (EDB) predicate is one whose contents can be directly manipulated by transactions that insert and remove records.
- An intensional (IDB) predicate (aka *view*) is defined by one or more Datalog rules. IDB predicates are maintained incrementally in response to changes made to EDB predicates.

A *primitive* is a function or relation that is calculated on demand. For example, the function  $\text{add}[x, y] = z$  is a primitive that computes  $z = x + y$ .

### 1.3.1 Key- and value-position

For a materialized predicate atom  $R(x_1, \dots, x_k)$  or  $F[x_1, \dots, x_k] = y$ , we say the variables  $x_1, \dots, x_k$  appear in *key-position*. (If  $F$  is a primitive operation, e.g.,  $\text{add}[x_1, x_2] = y$ , we do not count it as having key-position appearances of variables.)

A variable is deemed a *key* if it appears anywhere in key-position in the body of a rule; otherwise it is a *value*. A binding for the key variables of a rule uniquely determines the values. For example, in the expression  $F[x] = a, G[y] = b, r = a + b$ , the variables  $x, y$  are keys, and the variables  $a, b, r$  are values.

## 2 Maintaining head predicates

In this section we describe how to maintain head predicates as changes are made to satisfying assignments of the body.

### 2.1 Projection-free rules

A rule is *projection-free* if each atom in the head contains an appearance of every key variable. For example:

$$\forall x, y, z . R(x, y, z) \leftarrow A(x, y), B(y, z)$$

is projection-free, whereas:

$$\forall x, y . S(x, y) \leftarrow \exists z . A(x, y), B(y, z)$$

is not, because the key-variable  $z$  does not appear in the head atom  $S(x, y)$ .

Maintaining head predicates for projection-free rules is easy: we simply insert or remove records in response to the changes made to satisfying assignments of the body.

## 2.2 Rules with projection

For rules with projection we primarily use counting [5]. Consider the rule:

$$S(x, y) \leftarrow A(x, y), B(y, z)$$

We represent  $S$  by a predicate  $S[x, y] = \eta$ , where  $\eta$  is a *support count*: the number of satisfying assignments of the body producing  $(x, y)$ . (In the example rule, there might be several bindings of  $z$  for a given  $(x, y)$ .) Then, for  $\delta S(x, y, \Delta)$ , we respond to a  $\Delta = \text{INSERT}$  by incrementing  $\eta$ , and to a  $\Delta = \text{ERASE}$  by decrementing  $\eta$ . We use special data structure support (an *update-action*) that treats  $\eta$  as a reference count, so that a decrement of  $\eta$  resulting in  $\eta = 0$  causes the record to be deleted.

Functions appearing in rule heads are handled in a similar way: suppose the head predicate is  $F[s]=t$ . We maintain the head predicate as  $F[s]=(t, \eta)$ , where  $\eta$  is the support count. Given a set of deltas to apply, we order them so ERASE actions are applied first, to avoid issues with conflicting function values.

### 2.2.1 Short-circuit evaluation

In some cases we can avoid the use of reference counts by using *short-circuit evaluation*. For a rule such as:

$$S(x, y) \leftarrow A(x, y), B(y, z)$$

it is helpful to explicitly insert a quantifier for  $z$ :

$$S(x, y) \leftarrow \exists z. A(x, y), B(y, z)$$

Suppose the key order chosen by the optimizer is  $[x, y, z]$ . Given particular  $x, y$ , it is obviously of little use to enumerate all possible satisfying assignments for  $z$ . We can instead use short-circuit evaluation: as soon the first satisfying assignment for an  $(x, y)$  is produced, we can backtrack immediately without considering further assignments of  $z$ . In this case, the support count  $\eta$  is unnecessary. However, in some cases it might be more efficient to use a key order such as  $[z, y, x]$ , in which case short-circuit evaluation cannot be used. This decision is left to the query optimizer.

## 2.3 Aggregations

Our variant of Datalog supports aggregations such as sum, count, min, and max. For example, the following rule computes the total calories consumed by people from meals:

$$\begin{aligned} \text{CaloriesConsumed}[\text{person}] = \text{totcal} \leftarrow \\ \text{agg} \ll \text{totcal} = \text{sum}(\text{cal}) \gg \\ \text{ate}(\text{person}, \text{meal}), \text{caloriesOf}[\text{meal}] = \text{cal}. \end{aligned}$$

### 2.3.1 Aggregations: count

Count aggregations can be handled by using the same mechanism used for support counts of rules with projections. For example:

$$\begin{aligned} \text{outdegree}[x] = d \leftarrow \\ \text{agg} \ll d = \text{count}() \gg \\ E(x, y). \end{aligned}$$

can be implemented using a head predicate  $\text{outdegree}[x] = d$ , where  $d$  is a support count as described above. The support count is incremented and decremented in response to changes in satisfying assignments of the rule body: if a new satisfying assignment  $E(x, y)$  is found, then  $d$  is incremented; if a satisfying assignment is removed, then  $d$  is decremented, and the record is removed from  $\text{outdegree}$  when  $d = 0$ .

### 2.3.2 Aggregations: the Abelian group case

For sum aggregations over an Abelian group  $\langle G, +, -, 0 \rangle$ , where the operator  $+$  is associative and commutative, we can use an update-action that updates the aggregate by ‘adding’ the new value when  $\Delta = \text{INSERT}$ , and ‘adding’ the inverse (‘negative’) of the new value when  $\Delta = \text{ERASE}$ . We also employ a support count  $\eta$  to remove a record once no more satisfying assignments of the body contribute to it.

This style of aggregation can be used for sum aggregations over integers and fixed-precision data types.

### 2.3.3 Aggregations: the semigroup case

Min and max aggregations cannot be treated as Abelian group aggregations, since there is no inverse: i.e. no operation  $\cdot^{-1}$  such that  $\forall \alpha. \min(\alpha^{-1}, \alpha) = I$ , where  $I = -\infty$  is an identity element. We instead treat these as aggregations over a semigroup  $(M, \oplus)$ , where  $\oplus$  is a binary operator (e.g., min, max).

Consider the example aggregation:

$$\begin{aligned} A[x] = ms &\leftarrow \\ \text{agg} \ll ms = \max(s) \gg \\ D[x, y, z] = s. \end{aligned}$$

We use an intermediate predicate that supports scans, as described in Section 3.1. Each satisfying assignment of the body is inserted into this intermediate predicate by a special rule:

$$\begin{aligned} A_{\max\text{-scan}}[x, y, z] = s &\leftarrow \\ D[x, y, z] = s. \end{aligned}$$

The head predicate  $A[x] = ms$  is computed by performing scans on the  $A_{\max\text{-scan}}$  predicate, which we can write:

$$A[x] = \mathbf{Scan} (A_{\max\text{-scan}}, [x, -\infty, -\infty], [x, +\infty, +\infty])$$

That is, for each  $x$ ,  $A[x]$  is computed by taking a scan of all records in the interval from  $[x, -\infty, -\infty]$  to  $[x, +\infty, +\infty]$ , where  $-\infty, +\infty$  are the smallest/largest representable values of the datatype.

For each change in satisfying assignments of the body, we insert or remove records to/from the intermediate predicate  $A_{\max\text{-scan}}[x, y, z] = s$ , and then recompute whatever records of  $A[x] = ms$  could have changed. This lets us maintain the aggregation result in time  $O(\delta \log n)$ , where  $\delta$  is the number of changes in satisfying assignments of the body.

Note: we can reuse the intermediate predicate  $A_{\max\text{-scan}}[x, y, z] = s$  to provide aggregations at multiple levels of detail. For example, if we also wanted to know the maximum  $s$  for a given  $x, y$  pair, we could define a rule:

$$\begin{aligned} A'[x, y] = ms &\leftarrow \\ \text{agg} \ll ms = \max(s) \gg \\ D[x, y, z] = s. \end{aligned}$$

which could share the  $A_{\max\text{-scan}}[x, y, z] = s$  intermediate predicate with the rule calculating  $A[x] = ms$ :

$$A'[x, y] = \mathbf{Scan} (A_{\max\text{-scan}}, [x, y, -\infty], [x, y, +\infty])$$

String concatenation aggregations can also be handled using the semigroup approach; this can be made efficient by representing long strings using ropes [2].

### 2.3.4 Aggregations: floating-point sums

Floating-point sum aggregations are problematic because floating-point addition is not associative. The Abelian group approach described above would allow arbitrarily large errors to accumulate over time as the sum was maintained. The semigroup approach would produce answers that depended in subtle ways on the precise structure of the scan-tree (Section 3.1), due to nonassociativity; it would also have the undesirable requirement of storing all satisfying assignments of the rule body in an intermediate data structure.

The sensible alternative is to employ a head-predicate with an arbitrary-precision floating-point value, which lets us use the Abelian group approach. That is, for an aggregation such as:

$$\begin{aligned} F[x] &= tot \leftarrow \\ &\text{agg} \ll tot = total(v) \gg \\ G[x, y] &= v \end{aligned}$$

where  $v$  is a floating-point value, we use an intermediate predicate of the form  $F^*[x] = (tot^*, \eta)$ , where  $tot^*$  is represented using an arbitrary-precision type, and use the update technique mentioned in Section 2.3.2.

There is a useful trick that can be employed here to efficiently represent  $tot^*$ . Consider a floating-point sum  $S = \sum_{i \in I} s_i$ . Instead of representing  $S$  directly, we can instead represent the sum  $X + (\sum_{i \in I} s_i)$ , where  $X \in \mathbb{R}$  is a value such as:

$$X = \sum_{k=-512}^{512} 2^{4k}$$

The binary representation of  $X$  is e.g.:

$$10001000100010001000 \dots 1000100010001.000100010001000 \dots 1000100010001$$

We partition  $X$  into 52-bit segments, this being the number of mantissa bits in an IEEE 754 double-precision floating point number. We only store 52-bit segments of  $X + (\sum_{i \in I} s_i)$  that differ from the corresponding segment of  $X$ , representing each segment as a floating-point number. Since  $X$  has 1-bits at regular intervals, any borrowing required to accommodate a negative summand never requires increasing the Hamming distance between  $X$  and  $X + (\sum_{i \in I} s_i)$  by more than 4 bits. (Consider for example representing the sum of  $S = \{2^{500}, -1\}$ : with this representation we do not have to borrow from  $2^{500}$ , which would cause a run of 500 1's in the representation; instead we just borrow from  $2^4$ . We would store only two segments, the one containing  $2^{500}$  and the one containing  $2^0$ .)

To extract  $F[x] = tot$  from the intermediate predicate  $F^*[x] = (tot^*, \eta)$ , we use a rule of the form:

$$\begin{aligned} F[x] &= tot \leftarrow \\ F^*[x] &= (tot^*, \eta), \\ tot &= \text{toFloat}[tot^*]. \end{aligned}$$

The primitive `toFloat` is straightforward to implement: we identify the first bit-position where  $X + (\sum_{i \in I} s_i)$  differs from  $X$ . The value  $tot$  is positive if the first differing bit is zero, and negative if the first differing bit is one. We subtract  $X$ , and extract a 52-bit mantissa. Combined with an exponent and sign, this yields a double-precision floating-point quantity.

## 3 Algorithms & Data structures

### 3.1 Scans

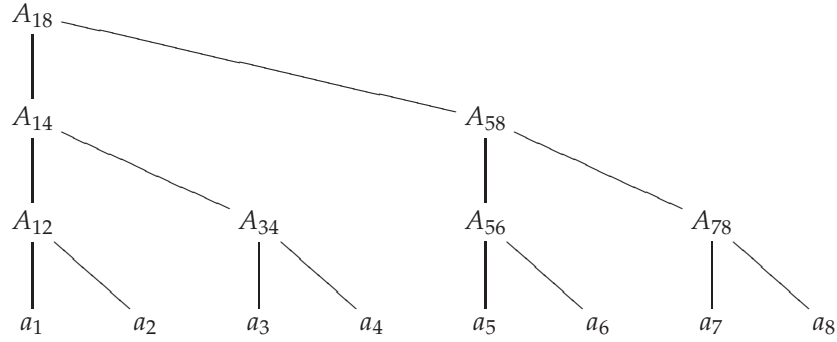
Scans (prefix-sums) are a handy formalism for aggregation-like operations [3]. We employ them for semigroup aggregations (Section 2.3.3), and also for implementing queries on sensitivity indices (Section 4.1).

Given an array  $A = [a_1, a_2, \dots, a_n]$  and an associative operator  $\oplus$ , the scan of  $A$  is just  $a_1 \oplus a_2 \oplus \dots \oplus a_n$ . If we choose  $\oplus$  to be addition, we get the sum; if we choose  $\oplus$  to be the max operator, we get the maximum element.



Suppose we want to calculate the aggregation over an arbitrary interval, i.e.  $a_i \oplus \dots \oplus a_j$  where  $1 \leq i \leq j \leq n$ . (For example, if the  $A$  array contained sales values for each day, we might want to aggregate sales over a specific month, rather than over all time.)

Associativity of the  $\oplus$  operator permits a simple data structure that can calculate the scan of any interval in  $O(\log n)$  time. Let  $A_{ij} = a_i \oplus \dots \oplus a_j$  be the aggregation over the elements  $a_i, \dots, a_j$ . We construct a binary tree (a *scan-tree*) with each leaf a single element  $a_i$ , and each internal node storing the  $\oplus$ -sum of its children:



For example, the left child of the root is:

$$\begin{aligned} A_{14} &= A_{12} \oplus A_{34} \\ &= (a_1 \oplus a_2) \oplus (a_3 \oplus a_4) \end{aligned}$$

To calculate the aggregation of an arbitrary interval we take the  $\oplus$ -sum of all subtrees contained entirely in the interval. For example:

$$\begin{aligned} a_1 \oplus \dots \oplus a_8 &= A_{18} \\ a_1 \oplus \dots \oplus a_3 &= A_{12} \oplus a_3 \\ a_2 \oplus \dots \oplus a_8 &= a_2 \oplus A_{34} \oplus A_{58} \\ a_3 \oplus \dots \oplus a_7 &= A_{34} \oplus A_{56} \oplus a_7 \\ a_4 \oplus \dots \oplus a_7 &= a_4 \oplus A_{56} \oplus a_7 \end{aligned}$$

For any interval  $a_i, \dots, a_j$ , we never need to sum more than  $2 \lceil \log_2 n \rceil = O(\log n)$  elements.

If a value changes, say  $a_5$  is changed to  $a'_5$ , we can update the scan-tree by simply recalculating all internal nodes on the path from  $a_5$  to the root ( $A_{56}, A_{58}, A_{18}$ ). This requires only  $O(\log n)$  operations.

For a concrete example, suppose we have a predicate  $\text{sales}[\text{region}, \text{store}] = \text{tot}$  giving the total sales for each store, and we wish to maintain the maximum sales of any store in each region:

$$\text{maxsales}[\text{region}] = \text{maxtot} \leftarrow \text{agg} \left\langle \begin{array}{l} \text{maxtot} = \max(\text{tot}) \\ \text{sales}[\text{region}, \text{store}] = \text{tot} \end{array} \right\rangle$$

Shown below is a scan-tree for calculating the max-aggregation of a sales predicate with 16 records. The first three columns give the (region, store, tot) records, and the scan-tree is drawn to the right. The records relevant for region

2 have been highlighted:

region	store	sales[region,store]
1	1	1000.00
1	2	1500.00
1	3	7300.00
1	4	8000.00
1	5	15000.00
2	6	2900.00
2	7	3500.00
2	8	1440.00
2	9	3300.00
2	10	1245.00
2	11	7024.00
2	12	5510.00
2	13	9000.00
3	14	325.00
3	15	4000.00
3	16	5300.00

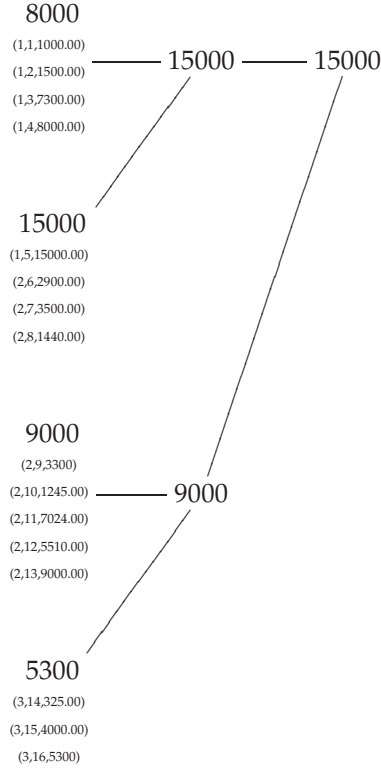
To calculate the maximum sales for region 2, we can just take the max of all subtrees for region 2:

```
maxsales[2] = max(2900.00, 3500.00, 7024.00, 9000.00)
               = 9000.00
```

### 3.1.1 Adapting scan-trees for paged data structures

In practice, we use a Btree-like data structure where leaf and index pages are augmented with scan information. Each leaf page is augmented with a ScanTree data structure that uses approximately 5% of the available space. It maintains a scan-tree for the records stored on the leaf page. This scan-tree uses a binary tree structure, but each scan-tree-leaf might aggregate a dozen or so records. For example, the sales data might be represented on a

Btree-like leaf page by this scan-tree:



The records being aggregated (shown in small font) are not actually stored in the scan-tree.

Each leaf in the scan-tree aggregates a variable number of records, so that insertions and deletions can be handled efficiently. The tree is occasionally rebalanced to ensure no child aggregates more than twice as many records as its sibling. This permits the scan tree to be updated in  $O(\log B)$  amortized time (where  $B$  is the Btree leaf page record capacity) in response to a record insert/update/delete.

For Btree-style index pages, we augment each record with an extra field containing scan information. On a Btree index page, records are typically of the form  $(key_1, \dots, key_n; pageid)$ , where  $pageid$  is the page number of a next-level leaf or index page. We add an additional scan-related field, so records are of the form  $(key_1, \dots, key_n; scan, pageid)$ , where  $scan$  aggregates all records in the subBtree reachable at  $pageid$ . In addition, each index page gets a ScanTree (taking approx. 5% of the page space) that aggregates the  $scan$  elements of the index-page records. This approach allows us to calculate the scan of any interval in an arbitrarily large predicate in  $O(\log n)$  time, where  $n$  is the number of records.

### 3.1.2 Efficient iteration of complements

Suppose we have a set  $S \subseteq T$ , and we wish to iterate the complement  $T \setminus S$ . This can be done efficiently using representations for  $S$  and  $T$  that include a scan-tree for a 'count' aggregation. Such a scan tree lets us count the number of records in an interval  $[k_1, k_2]$  in  $O(\log n)$  time, where  $k_1, k_2$  are keys (or key-tuples).

To iterate the complement, we can employ the principle that if a given key interval  $[k_1, k_2]$  contains the same number of records in  $S$  and  $T$ , then the complement  $T \setminus S$  is empty in that interval. This reduces the cost of iterating the complement to  $O(|T \setminus S| \cdot \log n)$ , a useful improvement for sparse complements. (The naive approach of iterating  $T$  and doing lookups in  $S$  would require  $O(|T|)$  time.)

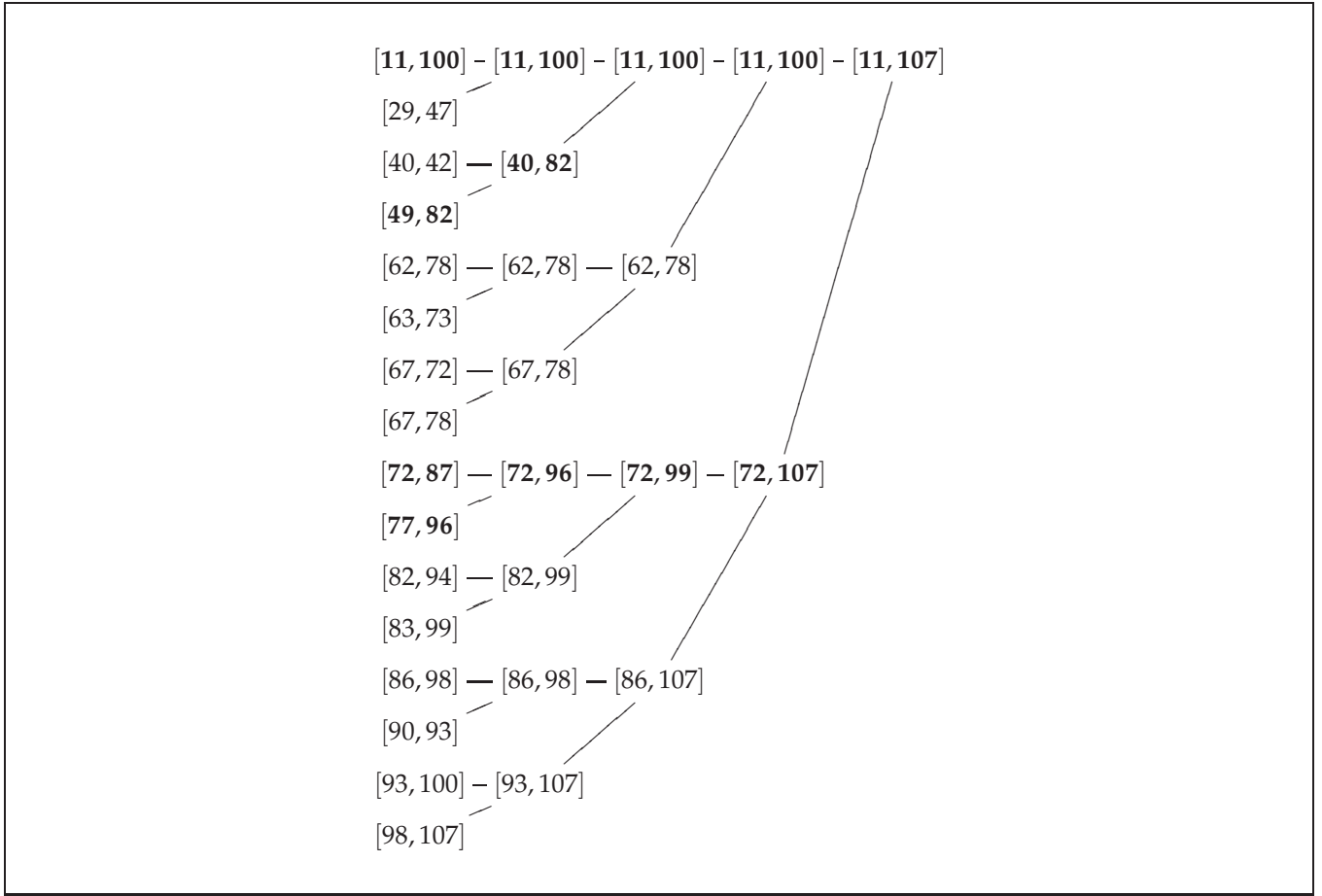


Fig. 2: An interval-tree. Nodes whose intervals contain  $x = 80$  are highlighted.

### 3.2 Interval trees

We use scan trees to implement interval trees, used to represent *sensitivity indices* in our maintenance algorithm (Section 4.1).

A simple interval tree stores a set of intervals  $I$ , with each interval of the form  $[a, b]$  where  $a, b \in K$  and  $K$  is some scalar key type. An interval query finds the set of intervals containing some key  $x$  of interest, i.e.

$$\text{IntervalQuery}(x) = \{[a, b] \in I : x \in [a, b]\}$$

For example, we might have

$$I = \{[2, 10], [3, 7], [5, 15], [6, 9]\}$$

and in response to the query ‘What intervals contain 10?’ it would produce  $\{[2, 10], [5, 15]\}$ .

To implement an interval tree, we can use a scan-tree, where each internal node of the scan tree has a pair  $[a, b]$ , where  $a$  is the min of the interval starts, and  $b$  is the max of the interval ends.

Figure 2 shows an example. The records are in the first column (integer intervals), and the scan-tree to the right. To find all intervals containing a particular number  $x$ , we start at the root and recursively descend to each child, backtracking when the scan-interval does not contain  $x$ . The nodes whose interval contain  $x = 80$  are highlighted above. The result set produced for  $x = 80$  is  $\{[11, 100], [49, 82], [72, 87], [77, 96]\}$ .

Interval trees produce the set of containing intervals for a value  $x$  in time  $O((m + 1) \log n)$ , where  $m$  is the number of matching intervals and  $n$  is the total number of intervals.

To adapt interval trees for paged data structures, we use a Btree augmented for scans, configured for a max-scan

on the endpoint of each interval. Since Btree-type data structures are ordered by key, and index pages store the least key of their subtrees, Btrees have a built-in min-aggregation for the startpoint of each interval.

For general sensitivity indices (Section 4.1), we use predicates with records of the form

$$\text{SensIndex}(\alpha_1, \alpha_2, \dots, \alpha_m, a, b, \gamma_1, \dots, \gamma_k)$$

This is understood to represent an interval of *tuples* beginning at  $[\alpha_1, \alpha_2, \dots, \alpha_m, a]$  and ending at  $[\alpha_1, \alpha_2, \dots, \alpha_m, b]$ . The  $\gamma_1, \dots, \gamma_k$  contain supplemental information described later.

### 3.3 Delta-iterators

Our current implementation of paged data structures use copy-on-write page-level versioning. This allows us to iterate through the difference between two consecutive versions of a predicate in  $O(\delta \log n)$  time, where  $\delta$  is the number of changes made between the two versions, and  $n$  is the maximum record count of the two versions. This is done by iterating through the two versions simultaneously, and skipping any subtrees common to both versions.<sup>1</sup>

## 4 Maintaining rule bodies

We now describe our maintenance algorithm for rule bodies. Recall the example:

$$F(x, y) \leftarrow G(x, z), H(y, z), I(x, y, z).$$

We wish to maintain  $F(x, y)$  given updated versions of the body predicates  $G', H', I'$ . We evaluate a maintenance rule of the form:

$$\delta F(x, y, \Delta) \leftarrow (\text{Body}[G, H, I] \cdots \text{Body}[G', H', I'])(x, y, z, \Delta), \\ \text{ChangeOracle}(x, y, z).$$

where  $(\text{Body}[G, H, I] \cdots \text{Body}[G', H', I'])(x, y, z, \Delta)$  tabulates changes in satisfying assignments of the rule body, and  $\text{ChangeOracle}(x, y, z)$  restricts evaluation to regions of the  $(x, y, z)$  tuple space where changes may occur. Roughly speaking, if you assert a new fact, the change-oracle tells you where it *could* be used; if you retract a fact, the oracle tells you where it *was* used. The use of the change-oracle is crucial to efficiency.

The  $\text{ChangeOracle}$  predicate is the essential heart of our maintenance algorithm. During initial full-evaluation of the rule for  $F$ , we build indices that note how changes to the predicates  $G, H, I$  might affect evaluation. To produce the  $\text{ChangeOracle}$  predicate, we use the differences between the body predicates  $(G \cdots G')$ ,  $(H \cdots H')$  and  $(I \cdots I')$  and these indices to produce the  $\text{ChangeOracle}$  predicate. Doing this efficiently requires some special algorithms and data structures described in Section 3.

Using delta-iterators (Section 3.3), we can efficiently enumerate the changes to the body predicates; let:

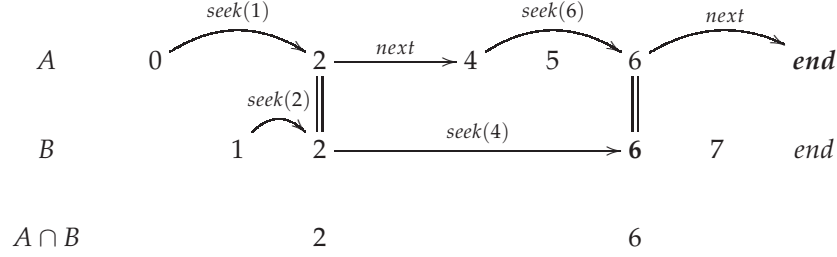
$$\begin{aligned} \delta G(x, z, \Delta) &= (G \cdots G')(x, z, \Delta) \\ \delta H(y, z, \Delta) &= (H \cdots H')(y, z, \Delta) \\ \delta I(x, y, z, \Delta) &= (I \cdots I')(x, y, z, \Delta) \end{aligned}$$

To construct the change-oracle, we need to determine what portions of the  $(x, y, z)$  tuple-space might need to be revisited, given the changes  $\delta G$ ,  $\delta H$ , and  $\delta I$ . For this we use *sensitivity indices* that record how the rule evaluation is sensitive to changes in the body predicates.

<sup>1</sup> A more sophisticated planned data structure, *cascading trees*, does versioning in a way that minimizes the number of pages altered. With cascading trees, the number of pages that must be examined for delta-iteration is  $O(\delta B^{-1/2} \log \delta)$ , where  $B$  is the average leaf-page capacity. In practice, this means that e.g. 50 changes, even to widely scattered keys, will usually be concentrated on a single page.

## 4.1 Sensitivity indices

In our Datalog system, queries are evaluated using the leapfrog triejoin algorithm (LFTJ) [6]. Here is an example leapfrog join for  $A(x), B(x)$ , where  $A = \{0, 2, 4, 5, 6\}$  and  $B = \{1, 2, 6, 7\}$ :



The leapfrog join begins by positioning an iterator at the start of each predicate, then repeatedly applying these rules (demonstrated by the arrows in the above diagram):

- If either iterator is at end, then stop.
- If both iterators are positioned at the same key, emit this key. Then increment one iterator.
- Otherwise, take the iterator positioned at the lesser key, and do a seek-lub to the key at which the other iterator is positioned.

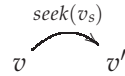
We count the trace as the operations performed on the iterators, and their result (e.g. one step in the above would be ‘seek(6) from  $x=4$  to  $x=6$  on iterator A’).

LFTJ can handle most  $\exists_1$  queries, but at the lowest level they are implemented in terms of trie-iterator operations such as *open()*, *up()*, *next()*, and *seek\_lub()*. So, the approach we are about to describe for maintaining  $A(x), B(x)$  extends naturally to more complex queries.

We want to know: what changes to  $A, B$  might cause changes to the trace? For example, if we inserted  $B(5)$ , this would change the trace, since the *seek(4)* arrow from  $x = 2$  to  $x = 6$  in B would change to land on  $x = 5$ . However, if we inserted  $B(3)$ , this would not change the trace, because the *seek(4)* arrow is seeking a least upper bound for 4; the trace is not *sensitive* to changes in B at  $x = 3$ .

The rules for trace sensitivity of a unary predicate  $D(x)$  are straightforward:

1. Seeks:



If the iterator for predicate D is positioned at key  $v$ , and a *seek\_lub( $v_s$ )* operation is performed so the iterator is then positioned at  $v'$ , then the trace is sensitive to changes in D in the interval  $[v_s, v']$ . (It is not sensitive to changes in  $(v, v_s)$ , because the seek operation finds a least upper bound for  $v_s$ .)

2. Increment:



If the iterator for D is positioned at key  $v$ , and an increment (*next*) is performed so the iterator is then positioned at  $v'$ , then the trace is sensitive to changes to D in the interval  $[v, v']$ .

3. If the iterator for predicate D is opened at position  $v$  (i.e. the first record is  $v$ ), then the trace is sensitive to changes in D in the interval  $(-\infty, v]$ .

For the above  $A(x), B(x)$  example, the sensitivities are:

$$A_{sens} = \{[-\infty, 0], [1, 2], [2, 4], [6, 6], [6, +\infty]\}$$

$$B_{sens} = \{[-\infty, 1], [2, 2], [4, 6]\}$$

Given changes  $\delta A, \delta B$ , we collect intervals where the predicate is sensitive *and* a change has occurred there:

$$\begin{aligned} A_{co}([x_1, x_2]) &\leftarrow \delta A(x, \Delta), x \in [x_1, x_2], A_{sens}([x_1, x_2]). \\ B_{co}([x_1, x_2]) &\leftarrow \delta B(x, \Delta), x \in [x_1, x_2], B_{sens}([x_1, x_2]). \end{aligned}$$

To evaluate the above rules for  $A_{co}, B_{co}$  efficiently, we use an interval-tree representation for  $A_{sens}$  and  $B_{sens}$  (Section 3.2).

We can then define the change-oracle as:

$$\text{ChangeOracle}(x) \equiv x \in [x_1, x_2], (A_{co}([x_1, x_2]); B_{co}([x_1, x_2])).$$

(In our notation, the semicolon indicates a disjunction.) For efficiency, we treat  $\text{ChangeOracle}(x)$  as a nonmaterialized predicate: during evaluation of the maintenance rule, the iterator for  $\text{ChangeOracle}(x)$  is internally manipulating the  $A_{co}$  and  $B_{co}$  predicates to present the contents of  $\text{ChangeOracle}(x)$ , without explicitly expanding the intervals into individual elements.

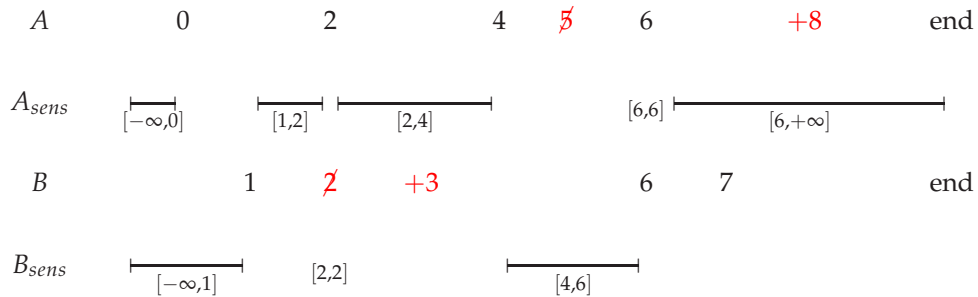
A note on the maintenance cycle: after each matching interval in  $A_{sens}$  is found, we remove it from  $A_{sens}$ ; this guarantees that the cost of evaluating the  $A_{co}$  rule is  $O((|\delta A| + |A_{co}|) \log n)$ , i.e. proportional to the number of changes and  $A_{co}$ -results. The  $\log n$  factor reflects the btree heights; a sharper estimate would be to take  $n = \max(|\delta A|, |A_{sens}|)$ . When the maintenance rule is evaluated, we accumulate new sensitivity intervals to  $A_{sens}$  and  $B_{sens}$ , so we are ready for the next round of maintenance.

#### 4.1.1 Example of maintenance for a unary join

For a concrete example, suppose we have:

$$\begin{aligned} \delta A &= \{(5, \text{ERASE}), (8, \text{INSERT})\} \\ \delta B &= \{(2, \text{ERASE}), (3, \text{INSERT})\} \end{aligned}$$

This diagram shows the changes made to A and B, and the sensitivity intervals:



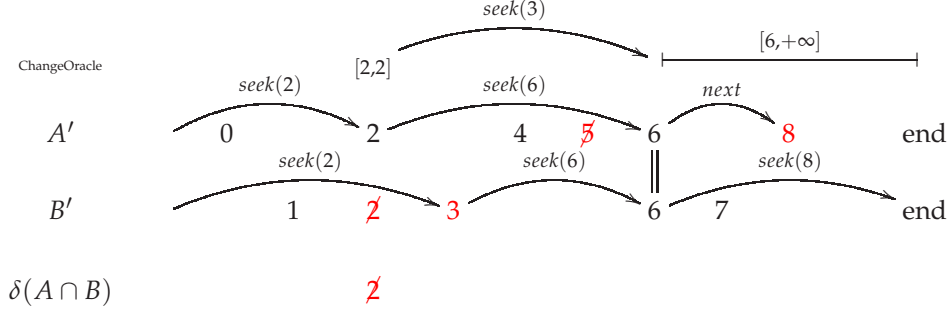
When we evaluate the rules for  $A_{co}, B_{co}$ , we find these contributions:

change	contributions to $A_{co}, B_{co}$
$A : (5, \text{ERASE})$	$\emptyset$
$A : (8, \text{INSERT})$	$\{[6, \text{end}]\}$
$B : (2, \text{ERASE})$	$\{[2, 2]\}$
$B : (3, \text{INSERT})$	$\emptyset$

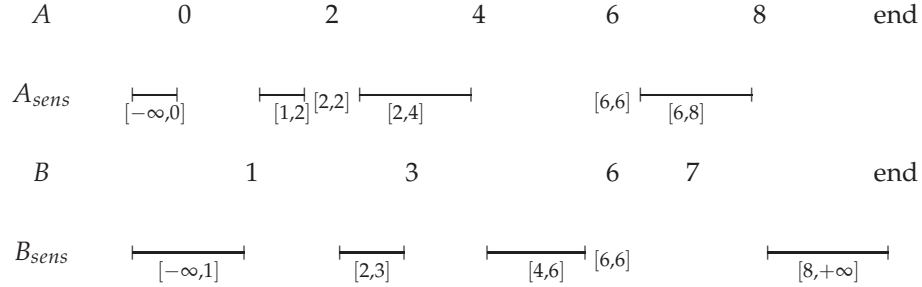
and so

$$\text{ChangeOracle}(x) = \bigcup \{[2, 2], [6, +\infty]\}$$

Maintenance: because of the ChangeOracle, we skip immediately to  $x = 2$ ; there we find that 2 is no longer in  $A \cap B$ . Then we skip to the start of the next interval in the ChangeOracle,  $x \in [6, +\infty]$ :



During evaluation of the maintenance rule, the sensitivity intervals are updated, so we are ready for the next round of maintenance: the intervals we examined because of the ChangeOracle are removed, and we insert new ones due to iterator operations as we evaluate the rule. The revised sensitivity intervals are:



## 4.2 Sensitivity indices for predicates with multiple arguments

For a trivial query like  $A(x), B(x)$ , sensitivity indices & the change-oracle are of marginal use; in fact our system would not use them for such a simple query. However, for complex queries these techniques make a tremendous difference.

Consider this example:

$$F(x, y) \leftarrow G(x, z), H(y, z), I(x, y, z), R(z).$$

Suppose the optimizer chooses the key-variable ordering  $[x, y, z]$ . If a fact is retracted from  $R(z)$ , the change-oracle lets us examine only those  $(x, y, z)$  tuples where that fact was used.

The approach to building the change-oracle for predicates with multiple key arguments generalizes that described in the previous section. First, a bit of background.

Recall that LFTJ evaluates rules using a ‘backtracking search through tuple space,’ which conceptually consists of nested leapfrog joins on unary predicates. We write  $G(x, \_)$  for projection, and  $G_x(y)$  for a curried version of  $G$  for a specific  $x$ . For instance, given  $G = \{(0, 10), (0, 20), (1, 30)\}$ , we would have:

$$\begin{aligned} G(x, \_) &= \{0, 1\} \\ G_0(y) &= \{10, 20\} \\ G_1(y) &= \{30\} \end{aligned}$$

(Note: we do not explicitly construct these projections and curried versions; this is just for exposition.)

The LFTJ algorithm does a backtracking search through the  $[x, y, z]$  space, first seeking a binding for  $x$ , then proceeding to a binding for  $y$  once  $x$  is found, etc. Conceptually, the three nested queries used are:



1.  $G(x, \_), I(x, \_)$
2.  $H(y, \_), I_x(y, \_)$
3.  $G_x(z), H_y(z), I_{xy}(z), R(z)$

When we evaluate the query, we record sensitivity information much as described earlier for unary predicates. However, we also record information about the bindings of other key-variables. The sensitivity predicate for  $R$ , for instance would have the form  $R_{sens,z}([z_1, z_2], x, y)$ . If a fact is removed from  $R$ , we can quickly determine the  $(x, y, z)$  bindings where that fact was used; if a fact is added, we can quickly determine where it could be used.

The sensitivity predicate for  $H_y(z)$  illustrates the general form:

$$H_{sens,z}(\underbrace{y}_{(1)}, \underbrace{[z_1, z_2]}_{(2)}, \underbrace{x}_{(3)})$$

In position (1) we have variables that precede  $z$  in the argument list for  $H$ ; in position (2) we have the sensitivity interval for  $z$ ; in position (3) we have key-variables that are bound before  $z$  but do not appear in the argument-list for  $H$ .

So, *conceptually*, we would have these sensitivity predicates:

$$\begin{aligned} &G_{sens,x}([x_1, x_2]) \\ &G_{sens,z}(x, [z_1, z_2], y) \\ &H_{sens,y}([y_1, y_2], x) \\ &H_{sens,z}(y, [z_1, z_2], x) \\ &I_{sens,x}([x_1, x_2]) \\ &I_{sens,y}(x, [y_1, y_2]) \\ &I_{sens,z}(x, y, [z_1, z_2]) \\ &R_{sens,z}([z_1, z_2], x, y) \end{aligned}$$

In practice we can drop sensitivity indices where the key-arguments of the atom form a prefix of the key-ordering chosen by the optimizer. For example, our implementation would not bother creating sensitivity indices for  $I$ , since its arguments match the chosen key order  $[x, y, z]$ ; it would also not create the index  $G_{sens,x}([x_1, x_2])$ , since  $(x)$  is also a prefix of  $[x, y, z]$ .

#### 4.2.1 Tree surgery operations

The delta-iterator described in Section 3.3 lets us efficiently enumerate the changed records between two consecutive versions of a predicate. For building the change oracle, we need finer information, namely, changes made to the *trie* presentation of the predicate. We call such changes *tree surgery operations*. Tree surgery operations consist of either inserting or removing branches.

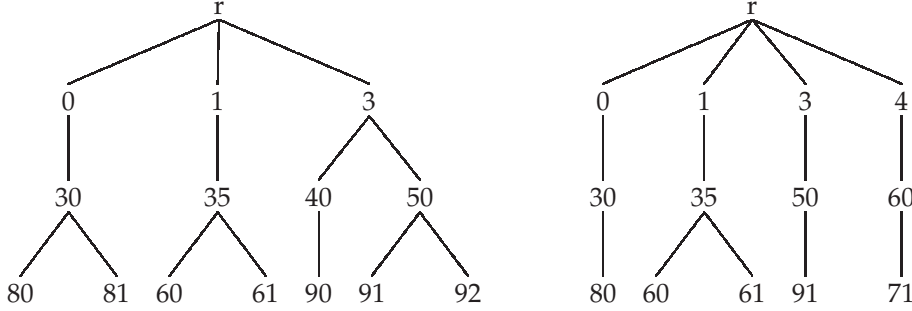
For example, consider these two versions of a predicate  $A(x, y, z)$ :

Version 1	Version 2
(0,30,80)	(0,30,80)
(0,30,81)	
(1,35,60)	(1,35,60)
(1,35,61)	(1,35,61)
(3,40,90)	
(3,50,91)	(3,50,91)
(3,50,92)	
	(4,60,71)

The delta-iterator would produce this stream of changes:

ERASE	0,30,81
ERASE	3,40,90
ERASE	3,50,92
INSERT	4,60,71

The trie presentations of the two versions are:



The tree surgery operations would be:

ERASE	0-30-81
ERASE	3-40-90
ERASE	3-40
ERASE	3-50-92
INSERT	4
INSERT	4-60
INSERT	4-60-71

It is reasonably straightforward and efficient to adapt a delta-iterator into an iterator of tree-surgery operations, with a little bookkeeping; we omit the details here.

#### 4.2.2 Matching tree surgery operations with sensitivity indices

Returning to our running example, recall that we have these two sensitivity indices for  $H(y, z)$ :

$$H_{sens,y}([y_1, y_2], x)$$

$$H_{sens,z}(y, [z_1, z_2], x)$$

We use a tree-surgery adaptor to get the changes made to the trie presentation of  $H$ , from the delta-iterator giving us the changes in  $H$ . Tree surgeries on  $H$  come in two forms: those that insert or remove vertices at depth 1, and those that insert or remove vertices at depth 2. We collect these surgeries by depth, writing  $\delta H^1(y, \Delta)$  and  $\delta H^2(y, z, \Delta)$  for depth-1 and depth-2 surgeries, respectively.

Trie surgery operations of depth 1 are matched with intervals in  $H_{sens,y}([y_1, y_2], x)$ , and trie surgery operations of depth 2 are matched to intervals in  $H_{sens,z}(y, [z_1, z_2], x)$ . The resulting change-oracle contributions we call  $H_{co,y}$  and  $H_{co,z}$ , and are defined by:

$$H_{co,y}(x, [y_1, y_2]) \leftarrow \delta H^1(y, \Delta), y \in [y_1, y_2], H_{sens,y}([y_1, y_2], x)$$

$$H_{co,z}(x, y, [z_1, z_2]) \leftarrow \delta H^2(y, z, \Delta), z \in [z_1, z_2], H_{sens,z}(y, [z_1, z_2], x)$$

As mentioned previously, this is implemented with interval trees and is very efficient—proportional (modulo  $\log n$ ) to the number of tree-surgery operations plus the number of matches to those operations in the sensitivity indices (Section 3.2). (Also recall that we remove matched intervals from the sensitivity indices.)

We define contributions to the change oracle from  $G$ ,  $I$ , and  $R$  similarly. Finally, we define the change oracle by

these (nonmaterialized) definitions:

$$\begin{aligned}
\text{ChangeOracle}(x, y, z) &\leftarrow \text{ChangeOracle}_1(x); \text{ChangeOracle}_2(x, y); \text{ChangeOracle}_3(x, y, z). \\
\text{ChangeOracle}_1(x) &\leftarrow (G_{co,x}([x_1, x_2]); I_{co,x}([x_1, x_2])), x \in [x_1, x_2]. \\
\text{ChangeOracle}_2(x, y) &\leftarrow (H_{co,y}(x, [y_1, y_2]); I_{co,y}(x, [y_1, y_2])), y \in [y_1, y_2]. \\
\text{ChangeOracle}_3(x, y, z) &\leftarrow (G_{co,z}(x, y, [z_1, z_2]); H_{co,z}(x, y, [z_1, z_2]); \\
&\quad I_{co,z}(x, y, [z_1, z_2]); R_{co,z}(x, y, [z_1, z_2])), z \in [z_1, z_2].
\end{aligned}$$

We can then maintain the rule, using the maintenance rule:

$$\begin{aligned}
\delta F(x, y, \Delta) &\leftarrow (\text{Body}[G, H, I] \cdots \text{Body}[G', H', I'])(x, y, z, \Delta), \\
&\quad \text{ChangeOracle}(x, y, z).
\end{aligned}$$

Recall that when evaluating the maintenance rule, we accumulate new intervals to the sensitivity indices, so we are ready for the next round of maintenance.

## References

- [1] Umut A. Acar, Guy E. Blelloch, Robert Harper, Jorge L. Vitter, and Shan Leung Maverick Woo. Dynamizing static algorithms, with applications to dynamic trees and history independence. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms, SODA '04*, pages 531–540, Philadelphia, PA, USA, 2004. Society for Industrial and Applied Mathematics.
- [2] Hans-Juergen Boehm, Russell R. Atkinson, and Michael F. Plass. Ropes: An alternative to strings. *Softw., Pract. Exper.*, 25(12):1315–1330, 1995.
- [3] Siddhartha Chatterjee, Guy E. Blelloch, and Marco Zagha. Scan primitives for vector computers. In *In Proceedings Supercomputing '90*, pages 666–675, 1990.
- [4] Songting Chen. *Efficient Incremental View Maintenance for Data Warehousing*. PhD thesis, Worcester Polytechnic Institute, 2005.
- [5] Ashish Gupta, Inderpal Singh Mumick, and V. S. Subrahmanian. Maintaining views incrementally. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD '93, Washington, DC, May 26–28, 1993*, volume 22(2) of *SIGMOD Record (ACM Special Interest Group on Management of Data)*, pages 157–166, pub-ACM:adr, 1993. ACM Press.
- [6] Todd L. Veldhuizen. Leapfrog triejoin: A worst-case optimal join algorithm. Technical Report LB1201, LogicBlox Inc., 2012. arXiv:1210.0481.