Joint work with:

[ICDT 24]

Idan Eldar

Benny Kimelfeld

Direct Access for Conjunctive Queries with Aggregation

Nofar Carmeli

TECHNION | The Henry and Marilyn Taub Faculty of Computer Science

Example

Goal: get a sense of how many views come from a contributor

Content

Contributor	Resource
Alice	CS101
Bob	CS101
Alice	Sophrology

Activity

Resource	Date	Views
CS101	01/01/23	4
CS101	02/01/23	125
Sophrology	01/01/23	26

Q(sum(views), contributor) ← content(contributor, resource), activity(resource, date, views)

1. Join

Contributor	Resource	Date	Views
Alice	CS101	01/01/23	4
Alice	CS101	02/01/23	125
Bob	CS101	01/01/23	4
Bob	CS101	02/01/23	125
Alice	Sophrology	01/01/23	26

2. Group by Contributor

Contributor	Resource	Date	Views
Alice Alice Alice	CS101 CS101 Sophrology	01/01/23 02/01/23 01/01/23	4 125 26
Bob Bob	CS101 CS101	01/01/23 02/01/23	4 125

3. Sum

Contributor	Views
Alice	155
Bob	129

4. Sort by Views

Views	Contributor
129	Bob
155	Alice

Example

Goal: get a sense of how many views come from a contributor

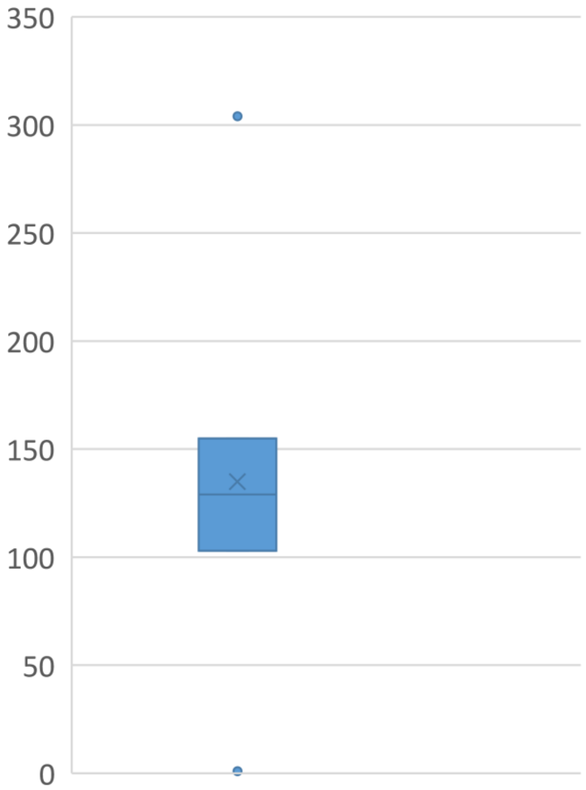
5. Get statistics

4. Sort by Views

Views	Contributor
1	Eve
103	Frank
117	Dave
129	Bob
136	Carol
155	Alice
304	George

350
max
200,250,300
3rd quartile
150
median
1st quartile
50,100
min
0

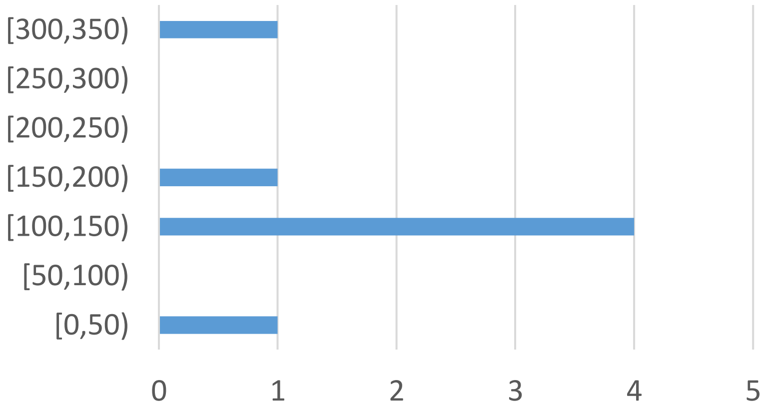
B) Boxplot



A) Median

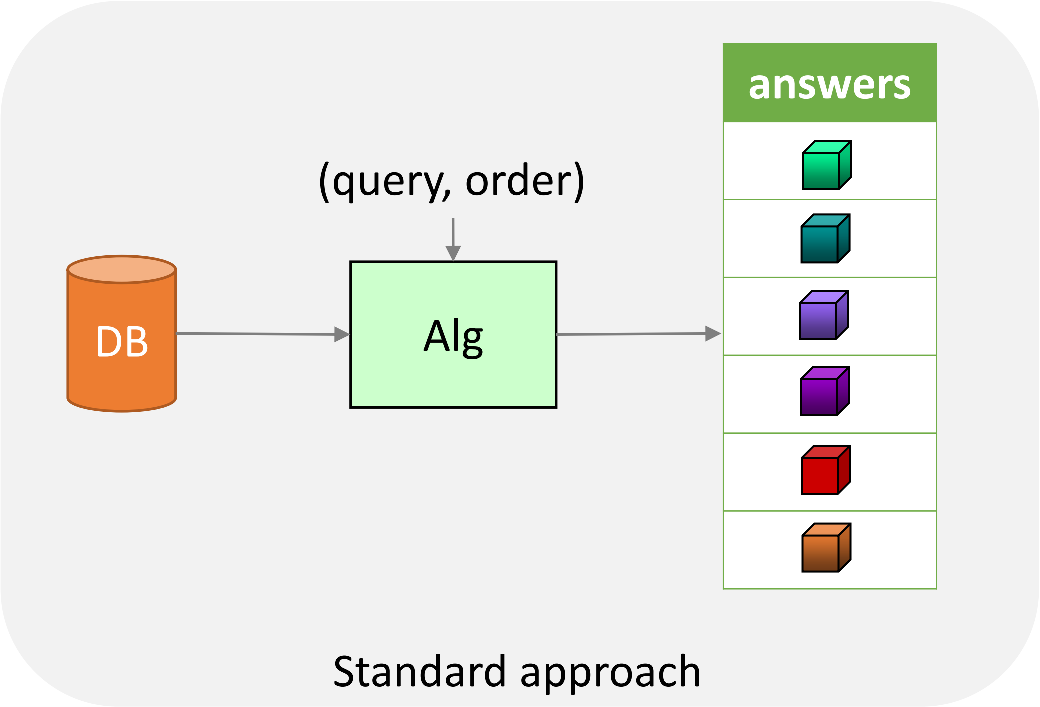
129 Bob

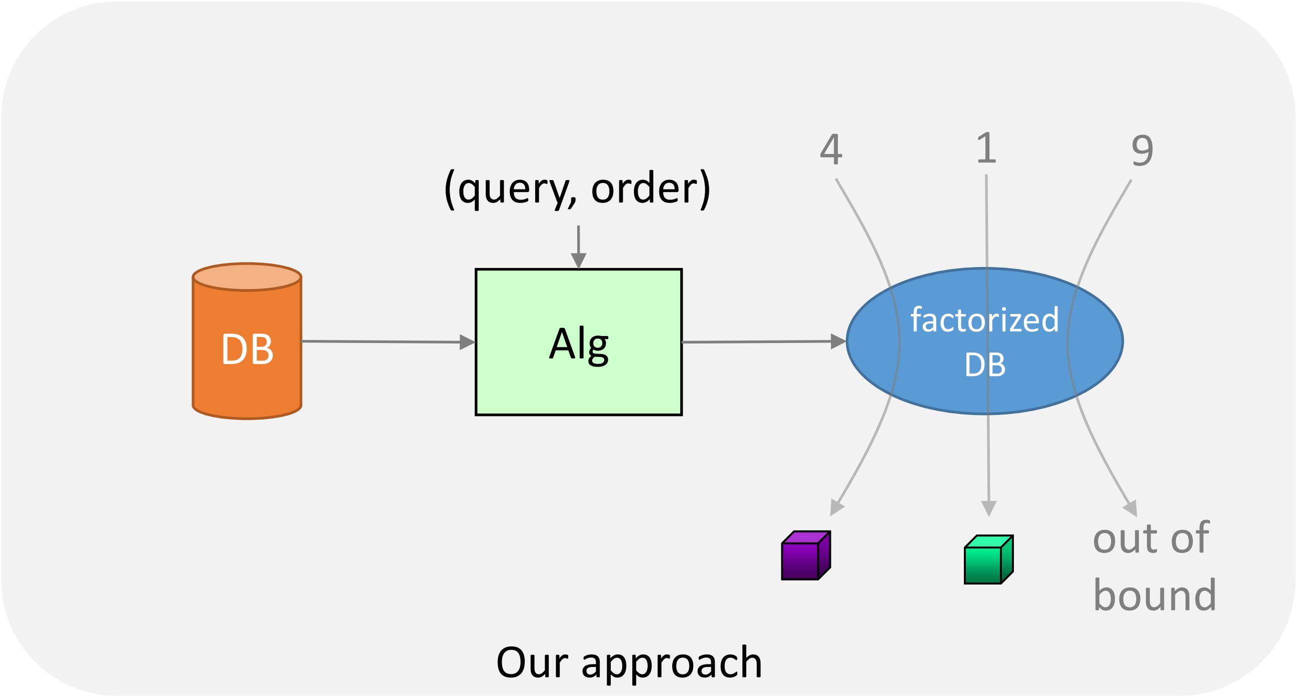
C) Histogram



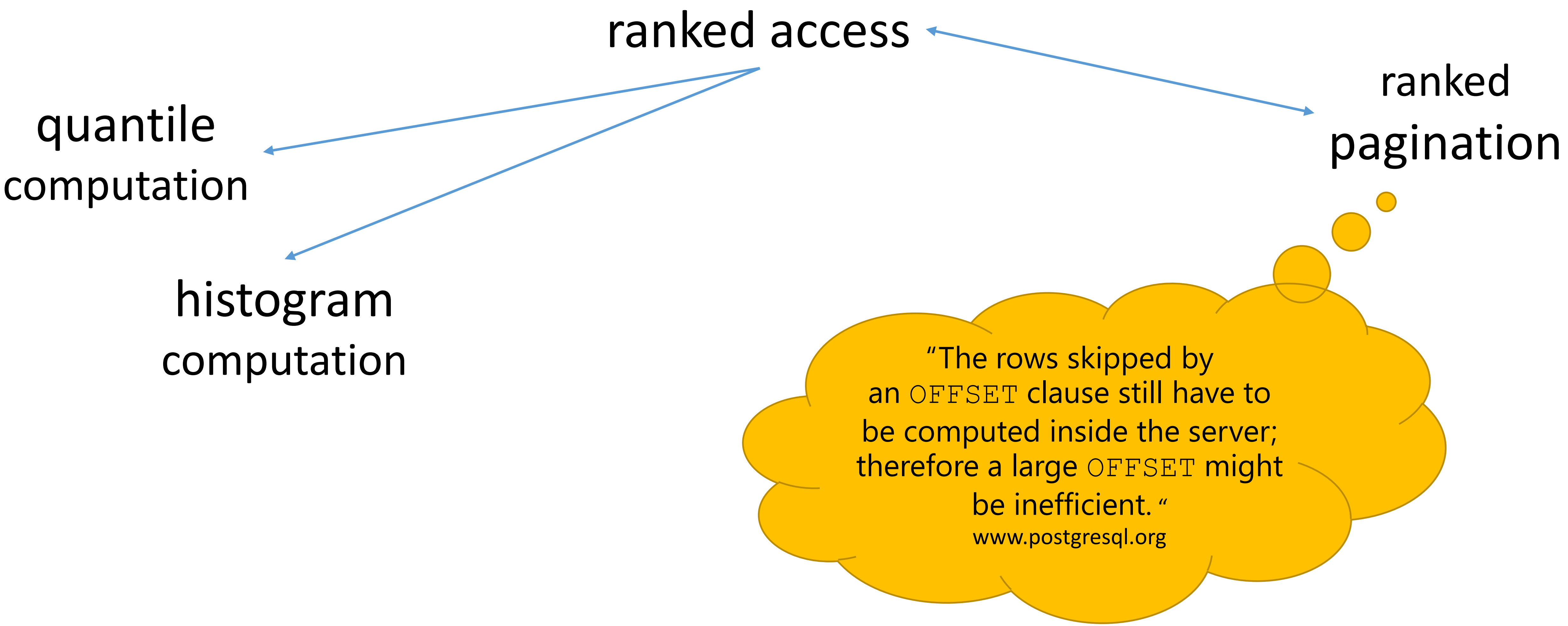
Definition: Ranked Direct Access

- Simulate a sorted array containing the answers
- Given i, returns the i^{th} answer or "out of bound".
- Ranked: user-specified order

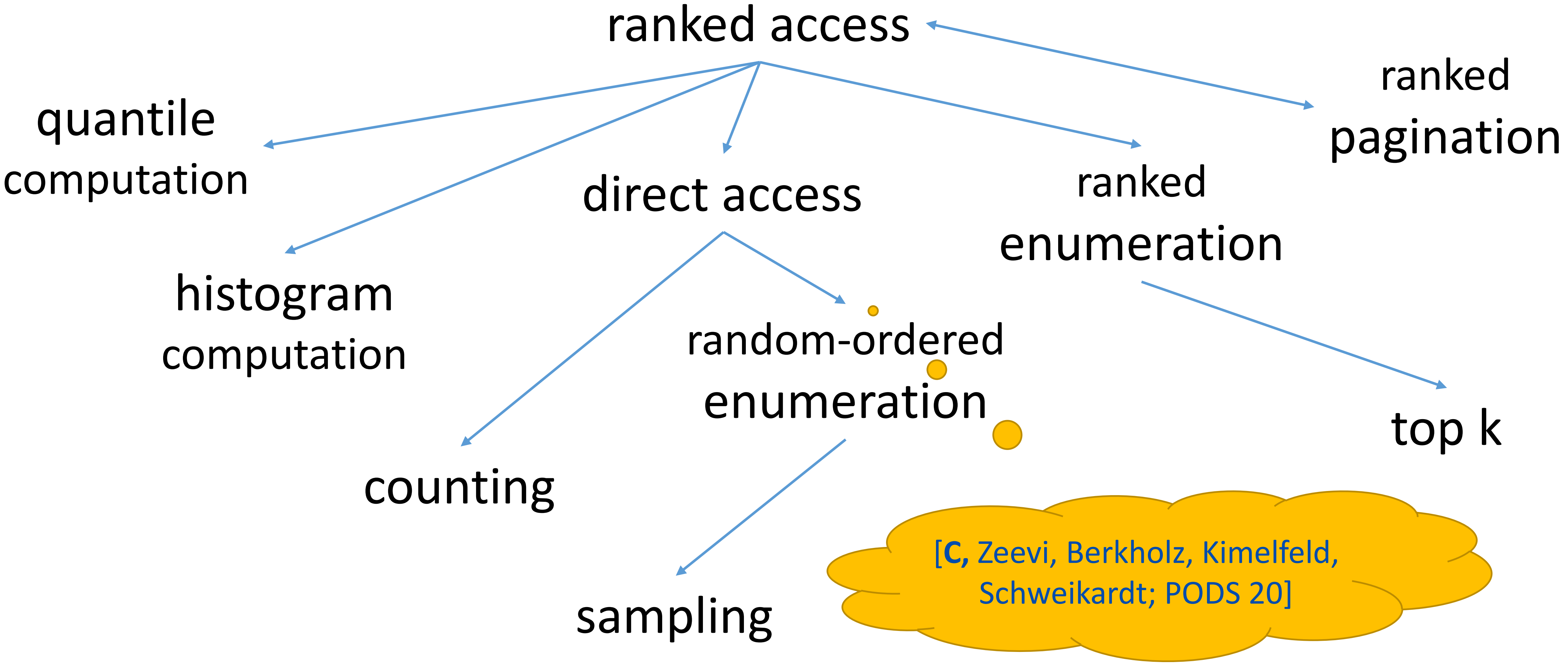




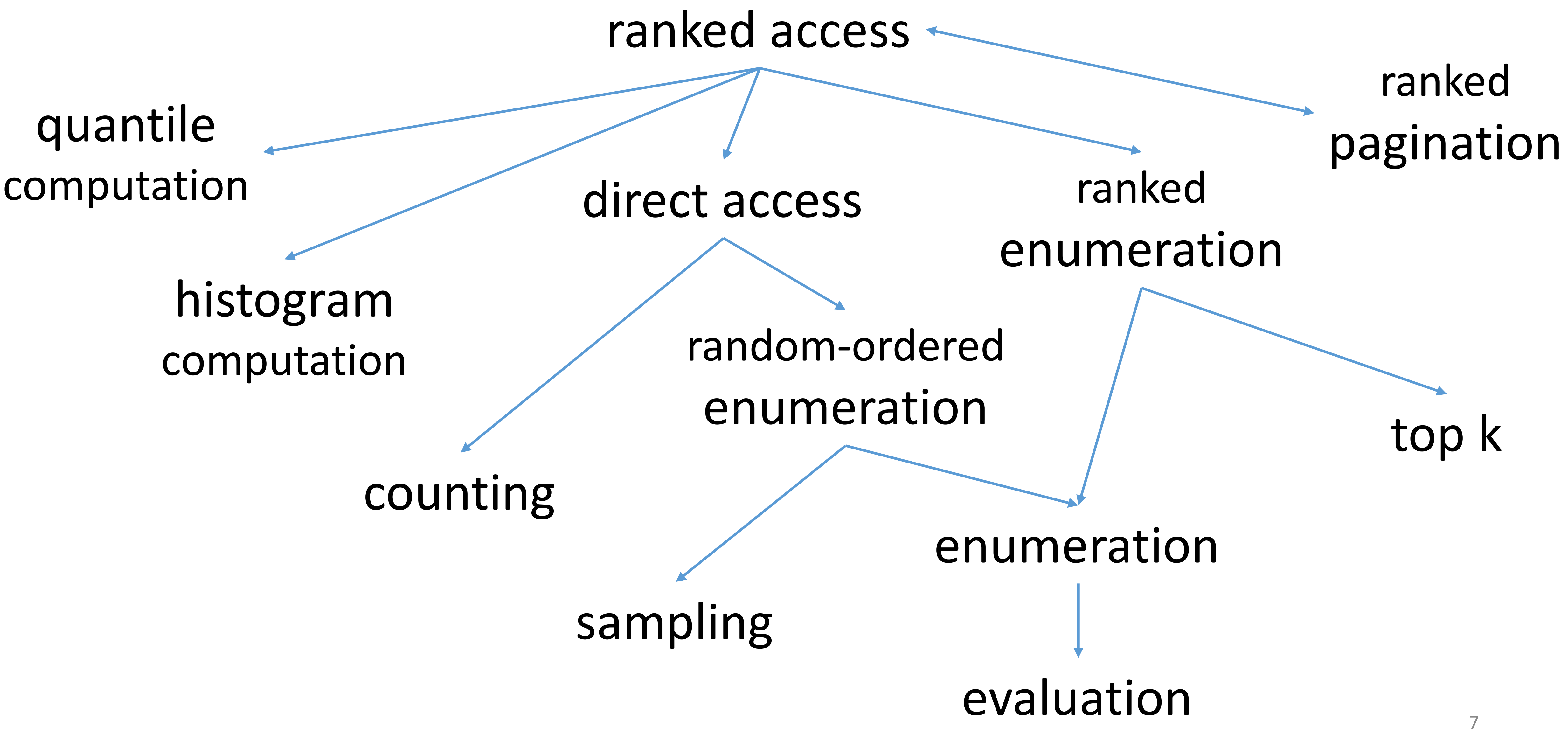
Overview of Tasks



Overview of Tasks



Overview of Tasks



Research question

When can we do ranked access with
(quasi)linear preprocessing and log access time?

Our focus: conjunctive queries with aggregation, lexicographic orders

Plan

- Motivation
- Dichotomy without aggregation
- Aggregation not affecting the order
  - Using annotations, the dichotomy still holds
- Aggregation affecting the order
  - Limited tractability using general annotations
  - Local annotations
    - In some cases (full query or idempotent semiring), equivalent to hardness of CQs with FDs
- Conclusion

Dichotomy for CQs (without aggregation)

[**C**, Tziavelis , Gatterbauer, Kimelfeld, Riedewald; PODS 21]

Given: conjunctive query Q, ordering L of free(Q),

lexicographic access in <loglinear,log>

$\Updownarrow$*

acyclic free-connex, no disruptive trio

* Lower bound requires:

sHyperclique hypothesis: $\forall k \geq 3$ the existence of a k-hyperclique in a $(k-1)$-uniform hypergraph cannot be decided in quasilinear time in the number of edges

sBMM hypothesis: Boolean matrices cannot be multiplied in quasilinear time in the number of the 1 entries

Definition: Free-Connex Acyclic

An acyclic CQ has a graph with:

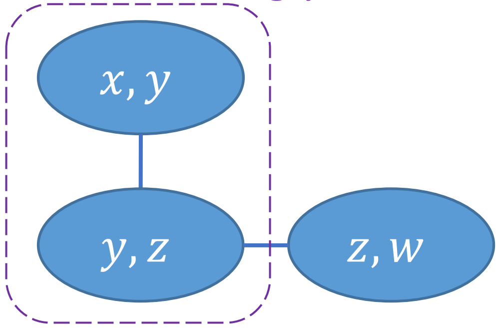
A free-connex CQ also requires:

1. a node for every atom

2. tree

3. for every variable:
the nodes containing it form a subtree

nodes containing y

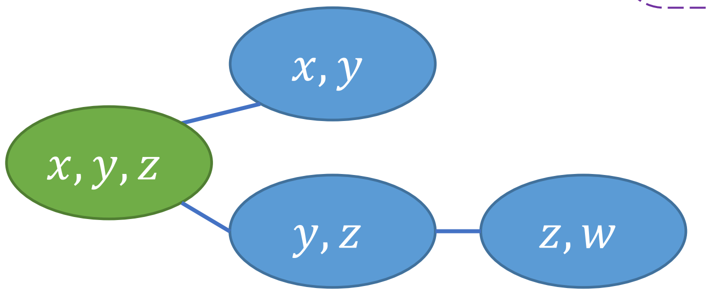


free − connex

acyclic

$$Q(x, y, z) \leftarrow R_1(x, y), R_2(y, z), R_3(z, w)$$

4. remains acyclic when introducing
an atom with the free variables



Dichotomy for CQs (without aggregation)

[**C**, Tziavelis , Gatterbauer, Kimelfeld, Riedewald; PODS 21]

Self-join-free assumption not required
[Bringmann, **C**, Mengel; 23]

Given: conjunctive query Q, ordering L of free(Q),

lexicographic access in <loglinear,log>
⇕*
acyclic free-connex, no disruptive trio

Disruptive Trio Definition

Examples

$Q_1(v_1, v_2, u) \leftarrow R(v_1, u), S(u, v_2)$
$Q_2(u, v_1, v_2) \leftarrow R(v_1, u), S(u, v_2)$

* Lower bound requires:
<u>sHyperclique hypothesis:</u> $\forall k \geq 3$ the existence of a k-hyperclique in a $(k-1)$-uniform hypergraph cannot be decided in quasilinear time in the number of edges
<u>sBMM hypothesis:</u> Boolean matrices cannot be multiplied in quasilinear time in the number of the 1 entries

Plan

- Motivation
- Dichotomy without aggregation
- Aggregation not affecting the order
  - Using annotations, the dichotomy still holds
- Aggregation affecting the order
  - Limited tractability using general annotations
  - Local annotations
    - In some cases (full query or idempotent semiring), equivalent to hardness of CQs with FDs
- Conclusion

Aggregation not affecting the order

- Approach: translate aggregates to semiring annotations.
- Example:

R

x	w
x_1	2
x_1	5
x_2	8

S

y
y_1
y_2

$Q(x, y, sum(w)) \leftarrow R(x,w), S(y)$

answers

x	y	$sum(w)$
x_1	y_1	$2+5$
x_1	y_2	$2+5$
x_2	y_1	8
x_2	y_2	8

1) translate

R

x	w	
x_1	2	2
x_1	5	5
x_2	8	8

S

y	
y_1	1
y_2	1

numerical semiring $(\mathbb{Q}, +, \cdot, 0, 1)$

$Q(x, y, \star) \leftarrow R(x,w), S(y)$

answers

x	y	
x_1	y_1	$(2+5)\cdot 1$
x_1	y_2	$(2+5)\cdot 1$
x_2	y_1	$8 \cdot 1$
x_2	y_2	$8 \cdot 1$

2) handle projections

R

x	
x_1	$2+5$
x_2	8

S

y	
y_1	1
y_2	1

3) Use access algorithm for CQs

The 2nd answer is: x_1 y_2

4) multiply annotations

The 2nd answer is:

x_1 y_2 $(2+5)\cdot 1$

Dichotomy for CQs with annotations last

Given: CQ⋆ $Q(\vec{x},\star)$

lexicographic access in <loglinear,log>

$\Updownarrow$*

acyclic free-connex, no disruptive trio

* Lower bound requires:

sHyperclique hypothesis: $\forall k \geq 3$ the existence of a k-hyperclique in a $(k-1)$-uniform hypergraph cannot be decided in quasilinear time in the number of edges

sBMM hypothesis: Boolean matrices cannot be multiplied in quasilinear time in the number of the 1 entries

Using Log-time Commutative Semirings

- Commutative semiring: $(\mathcal{K}, \oplus, \otimes, \bar{0}, \bar{1})$
  - $\mathcal{K}$ is a domain of elements
  - $(\mathcal{K}, \oplus, \bar{0})$ is a commutative monoid (“addition”)
    - $(a \oplus b) \oplus c = a \oplus (b \oplus c)$ (associative)
    - $a \oplus b = \mathrm{b} \oplus a$ (commutative)
    - $a \oplus \bar{0} = a$ ($\bar{0}$ neutral)
  - $(\mathcal{K}, \otimes, \bar{1})$ is a commutative monoid (“multiplication”)
  - $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$ (distributive)
  - $a \otimes \bar{0} = \bar{0}$
- In databases [Green, Karvounarakis, Tannen 2007]:
  - Each tuple is annotated with a semiring element
  - When joining tuples, multiply the annotations
  - When projecting, sum up the group’s annotation

Aggregations and Semirings

- Using log-time commutative semirings:
  - Sum: numerical semiring $(\mathbb{Q}, +, \cdot, 0, 1)$
  - Count: counting semiring $(\mathbb{N}, +, \cdot, 0, 1)$
  - Min: min-tropical semiring $(\mathbb{Q} \cup \{\infty\}, \min, +, \infty, 0)$
  - Max: max-tropical semiring $(\mathbb{Q} \cup \{-\infty\}, \max, +, -\infty, 0)$
- Average:
  - combine sum and count
- Count-Distinct:
  - No semiring translation
  - Harder than the others
    - $Q(x, \text{distinct}(z)) \leftarrow R(x, y), S(y, z)$ hard (assuming small-universe hitting set conjecture)
  - In case of log-size domain: use set semiring $(2^{\Omega}, \cup, \cap, \emptyset, \Omega)$

Small-universe Hitting Set Conjecture [Williams 15]:
Given two sets U and V of size N, each containing sets over $\{1,2,\ldots,d\}$, does U contains a set that shares an element with every set in V? Conjecture: it takes $N^{2-o(1)}$ time for every function $d = \omega(\log N)$.

Plan

- Motivation
- Dichotomy without aggregation
- Aggregation not affecting the order
  - Using annotations, the dichotomy still holds
- Aggregation affecting the order
  - Limited tractability using general annotations
  - Local annotations
    - In some cases (full query or idempotent semiring), equivalent to hardness of CQs with FDs
- Conclusion

Incorporating Aggregation in the Order

- Examples:
  - $Q_1(x, y, \star) \leftarrow R(x), S(y)$ easy (from dichotomy)
  - $Q_2(\star, x, y) \leftarrow R(x), S(y)$ hard (assuming 3SUM)
  - $Q_3(x, \star, y) \leftarrow R(x), S(y)$ easy (from sufficient condition)

<u>Sufficient condition:</u>
Consider a CQ⋆ $Q(\vec{x}, \star, \vec{z})$.
If every atom contains either all of $\vec{z}$ or none of $\vec{z}$,
and $Q'(\vec{x}, \vec{z})$ is acyclic free-connex with no disruptive trio, then*
lexicographic access in <loglinear,log> for $Q(\vec{x}, \star, \vec{z})$.

* Assuming ⊗-monotonicity.

<u>⊗-monotonicity:</u>
for every c, either $c \otimes a \preccurlyeq c \otimes b$ whenever $a \preccurlyeq b$, or $c \otimes b \preccurlyeq c \otimes a$ whenever $a \preccurlyeq b$.

<u>3SUM Conjecture:</u>
given a set of N elements from $\{-N^3, \ldots, N^3\}$, are there distinct elements a, b, c such that $a + b = c$? Conjecture: it takes $N^{2-o(1)}$ time.

Incorporating Aggregation in the Order

- Examples:
  - $Q_1(x, y, \star) \leftarrow R(x), S(y)$ easy (from dichotomy)
  - $Q_2(\star, x, y) \leftarrow R(x), S(y)$ hard (assuming 3SUM)
  - $Q_3(x, \star, y) \leftarrow R(x), S(y)$ easy (from sufficient condition)
  - $Q_4(sum(w), x, y) \leftarrow R(x, w), S(y)$ easy (locally annotated)
    - Translated to the hard $Q_2(\star, x, y) \leftarrow R(x), S(y)$
    - However, diverse annotation only in R
    - Equivalent in hardness to the easy $Q_4{}'(z, x, y) \leftarrow R(x, z), S(y)$ with the FD $x \rightarrow z$

Use FDs for more tractable cases
[C, Tziavelis, Gatterbauer, Kimelfeld, Riedewald; TODS 23]

R

x	w
x_1	2
x_1	5
x_2	8

S

y
y_1
y_2

$Q_4(sum(w), x, y) \leftarrow R(x, w), S(y)$

translate →

R

x	
x_1	7
x_2	8

S

y	
y_1	1
y_2	1

numerical semiring $(\mathbb{Q}, +, \cdot, 0, 1)$

$Q_2(\star, x, y) \leftarrow R(x), S(y)$

translate →

R

x	z
x_1	7
x_2	8

S

y
y_1
y_2

$Q_4{}'(z, x, y) \leftarrow R(x, z), S(y)$

Incorporating Aggregation in the Order

- Examples:
  - $Q_1(x, y, \star) \leftarrow R(x), S(y)$ easy (from dichotomy)
  - $Q_2(\star, x, y) \leftarrow R(x), S(y)$ hard (assuming 3SUM)
  - $Q_3(x, \star, y) \leftarrow R(x), S(y)$ easy (from sufficient condition)
  - $Q_4(sum(w), x, y) \leftarrow R(x, w), S(y)$ easy (locally annotated)
    - Translated to the hard $Q_2(\star, x, y) \leftarrow R(x), S(y)$
    - However, diverse annotation only in R
    - Equivalent in hardness to the easy $Q_4'(z, x, y) \leftarrow R(x, z), S(y)$ with the FD $x \rightarrow z$

Use FDs for more tractable cases [**C**, Tziavelis, Gatterbauer, Kimelfeld, Riedewald; TODS 23]

Full classification for local annotations in self-join-free case of:
full CQ⋆ or ⨁-idempotent semiring

Min
Max
Distinct (log domain)

⨁-idempotent: for every a, $a \oplus a = a$.

Incorporating Aggregation in the Order

- Examples:
  - $Q_1(x, y, \star) \leftarrow R(x), S(y)$ easy (from dichotomy)
  - $Q_2(\star, x, y) \leftarrow R(x), S(y)$ hard (assuming 3SUM)
  - $Q_3(x, \star, y) \leftarrow R(x), S(y)$ easy (from sufficient condition)
  - $Q_4(sum(w), x, y) \leftarrow R(x, w), S(y)$ easy (locally annotated)
    - Translated to the hard $Q_2(\star, x, y) \leftarrow R(x), S(y)$
    - However, diverse annotation only in R
    - Equivalent in hardness to the easy $Q_4'(z, x, y) \leftarrow R(x, z), S(y)$ with the FD $x \rightarrow z$
  - $Q_5(count(), x, y) \leftarrow R(x, w), S(y, z)$ easy (ad-hoc algorithm)

R

x	w
x_1	w_1
x_1	w_2
x_2	w_1

S

y	z
y_1	z_1
y_1	z_2
y_1	z_3

$Q_5(count(), x, y) \leftarrow R(x, w), S(y, z)$

translate

R

x	
x_1	2
x_2	1

S

y	
y_1	3

counting semiring $(\mathbb{N}, +, \cdot, 0, 1)$

$Q_2(\star, x, y) \leftarrow R(x), S(y)$

Conclusion

- Summary
  - Motivation
  - Dichotomy without aggregation
  - Aggregation not affecting the order
    - Using annotations, the dichotomy still holds
  - Aggregation affecting the order
    - Limited tractability using general annotations
    - Local annotations
      - In some cases (full query or idempotent semiring), equivalent to hardness of CQs with FDs
- Outlook
  - Open cases
  - Self-Joins
  - Time requirements for hard cases
    - Known for join queries [Bringmann, **C**, Mengel; PODS 22]
  - More complicated settings
    - Other orders
    - Other queries
    - Supporting updates