Answering Aggregate Queries with Ordered Direct Access

Idan Eldar

Joint Work: Nofar Carmeli, Benny Kimelfeld
Direct Access

**Standard Practice**
1. Run a full query computation
2. Construct an array-like structure
3. Retrieve a subset of the results

**Alternative Approach**
1. Construct a compact data structure supporting array-like access to the answers
2. Compute required results during access

database

full query results

\[ O(n) \]

\[ O(k) \]

\[ O(n^k) \]

\[ k \ll n \]
Target Complexity

**Standard Practice**
1. Run a full query computation
2. Construct an array-like structure
3. Retrieve a subset of the results

**Alternative Approach**
1. Construct a compact data structure supporting array-like access to the answers
2. Compute required results during access

\[
O(n^k) \quad \text{crsr.execute("query")} \quad \text{res = crsr.fetchall()} \quad \text{return res[start:start+10]} \quad O(n \log n) \quad O(\log n) \\
O(1) \quad \text{Read the data} \quad \text{Output a result}
\]

What complexity can we hope to achieve?

<table>
<thead>
<tr>
<th>Data</th>
<th>(O(n))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td>(O(k))</td>
</tr>
</tbody>
</table>

We denote this time complexity \(\langle \text{loglinear}, \log \rangle\)
Conjunctive Queries

• State-of-the-art results concern conjunctive queries (Select-Project-Join)

• Example:
  \[ Q(Hotel, City) : \neg \text{Flights}(Airline, City, Price), Hotels(Hotel, City) \]

• Conjunctive Queries:
  \[ Q(\bar{x}) : \neg \phi_1(\bar{x}, \bar{y}), \ldots, \phi_l(\bar{x}, \bar{y}) \]

• Known Dichotomy:
  [Bagan, Durand, Grandjean CSL'2007]
  [Brault-Baron 2013]
  [Carmeli et al. PODS2021]

  If \( Q \) is acyclic free-connex, and the order has no disruptive-trio,
  then direct access for \( Q \) is in \( \langle \text{loglinear}, \text{log} \rangle \)

  Otherwise, direct access is not in \( \langle \text{loglinear}, \text{log} \rangle \) *

* No self-joins, assuming BMM, HYPERCLIQUE conjectures
Aggregate Queries

• Conjunctive queries comprise only a subset of queries

• Example:

\[ Q(\text{Hotel, City, Min(Price)}) : \neg \text{Flights(Airline, City, Price), Hotels(Hotel, City)} \]

• General Form:

\[ Q(\vec{x}, \alpha(\vec{y}), \vec{z}) : \neg \phi_1(\vec{x}, \vec{y}, \vec{z}), ... , \phi_l(\vec{x}, \vec{y}, \vec{z}) \]

• Most common functions for \( \alpha \):
  • Min, Max, Sum, Avg, Count, Count Distinct (CountD)

• CQs are a particular case of ACQs
Aggregate Queries – Intractable

Non Acyclic Free-Connex ACQs*

Order contains a disruptive trio*

* No self-joins, assuming BMM, HYPERCLIQUE conjectures
Aggregate Queries Domain Split

Non Acyclic Free-Connex ACQs*

Aggregation is not part of the order

\[ Q(\text{Hotel, City, Min(Price)}) \]

Aggregation is part of the order

\[ Q(\text{Min(Price), Hotel, City}) \]

Order contains a disruptive trio*

* No self-joins, assuming BMM, HYPERCLIQUE conjectures
Aggregate Queries - Tractability

• Aggregation is not part of the order - $Q(\vec{x}, \alpha(\vec{y})): -\phi_1(\vec{x}, \vec{y}), ..., \phi_l(\vec{x}, \vec{y})$
• Acyclic free-connex ACQ $Q$
• No disruptive trio in $\vec{x}$

$\alpha$ is one of Min, Max, Sum, Count, Avg
Annotations

• In fact we prove an even stronger notion
• The CQs annotations framework allows tagging tuples with additional data and propagating it to the result

• Define the following:
  • Annotation domain - $K$
  • Projection propagation - $\oplus$
  • Join propagation - $\otimes$
  • Projection identity - $\overline{0}$
  • Join identity - $\overline{1}$

For every $k \in K$:

\[
\begin{align*}
  k \oplus \overline{0} &= k \\
  k \otimes \overline{1} &= k
\end{align*}
\]

[Green, Karvounarakis, Tannen 2007]
CQ* - Tractability

• Annotation is not part of the order - $Q(\vec{x}, *)$: $-\phi_1(\vec{x}, \vec{y})$, ..., $\phi_l(\vec{x}, \vec{y})$
• Acyclic free-connex CQ* $Q$
• No disruptive trio in $\vec{x}$

The $\oplus$, $\otimes$ operations of $(K, \oplus, \otimes, \overline{0}, \overline{1})$ can be performed in $O(\log n)$
CQ*s to ACQs

• Many ACQs can be solved using CQ*s with a proper semiring

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>(N, +, ·, 0, 1)</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>Sum</td>
<td>(Q, +, ·, 0, 1)</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>Max</td>
<td>(Q $\cup$ {$-\infty$}, max, +, $-\infty$, 0)</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>Min</td>
<td>(Q $\cup$ {$\infty$}, min, +, $\infty$, 0)</td>
<td>$O(1)$</td>
</tr>
</tbody>
</table>

The $\oplus$, $\otimes$ operations of $(K, \oplus, \otimes, 0, 1)$ can be performed in $O(\log n)$

$\alpha$ is one of Min, Max, Sum, Count, Avg
CQ*s Domain Split - Progress

Non Acyclic Free-Connex CQ*s*

$O(\log n)$ time $\otimes, \oplus$

Aggregation is part of the order

Order contains a disruptive trio*

* No self-joins, assuming BMM, HYPERCLIQUE conjectures
ACQs Domain Split - Progress

Non Acyclic Free-Connex ACQs*

\( \alpha \) is one of Min, Max, Sum, Count, Avg

Aggregation is part of the order

Count Distinct

Order contains a disruptive trio*

* No self-joins, assuming BMM, HYPERCLIQUE conjectures
ACQs - Tractability

• $Q(x_1, ..., \alpha(y), x_{i+1}, ..., x_k): -\phi_1(x, \bar{y}, \bar{z}), ..., \phi_l(x, \bar{y}, \bar{z})$

• Acyclic free-connex ACQ $Q$

• No disruptive trio in $\vec{x}$ when ignoring $\alpha(y)$

\[
\alpha \text{ is Count, Sum, Min or Max}
\]

Every atom $\phi_i$ contains either all or none of $x_{i+1}, ..., x_k$
CQ* s - Tractability

• $Q(x_1, ..., *, x_{i+1}, ..., x_k): -\phi_1(x, y), ..., \phi_l(x, y)$
• Acyclic free-connex CQ* $Q$
• No disruptive trio in $\vec{x}$ when ignoring $*$

$\otimes$ and $\oplus$ can be computed in $O(\log n)$

Every atom $\phi_i$ contains either all or none of $x_{i+1}, ..., x_k$

* The semiring must be $\otimes$-monotone:
  (All of ours are)
## CQ*s Domain Split - Progress

<table>
<thead>
<tr>
<th>Non Acyclic Free-Connex CQ<em>s</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>( O(\log n) ) time ( \otimes, \oplus )</td>
</tr>
<tr>
<td>Every atom ( \phi_i ) contains either all or none of ( x_{i+1}, \ldots, x_k )</td>
</tr>
</tbody>
</table>

| Order contains a disruptive trio* |

* No self-joins, assuming BMM, HYPERCLIQUE conjectures
ACQs Domain Split - Progress

Non Acyclic Free-Connex ACQs*

- $\alpha$ is one of Min, Max, Sum, Count, Avg
  
  Every atom $\phi_i$ contains either all or none of $x_{i+1}, \ldots, x_k$

- Count Distinct

- Order contains a disruptive trio*

* No self-joins, assuming BMM, HYPERCLIQUE conjectures
Intractability for CQ*s

• We can prove for our semirings a counter-example:

\[ Q_x(*, x, y): \neg R(x), S(y) \]

**Theorem:** Direct access for \( Q_x(*, x, y): \neg R(x), S(y) \) is not in \( \langle \text{loglinear, log} \rangle^* \)

(over these semirings)

<table>
<thead>
<tr>
<th>Semiring</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\mathbb{N}, +, \cdot, 0, 1))</td>
<td>Count</td>
</tr>
<tr>
<td>((\mathbb{Q}, +, \cdot, 0, 1))</td>
<td>Sum</td>
</tr>
<tr>
<td>((\mathbb{Q} \cup {-\infty}, \max, +, -\infty, 0))</td>
<td>Max</td>
</tr>
<tr>
<td>((\mathbb{Q} \cup {\infty}, \min, +, \infty, 0))</td>
<td>Min</td>
</tr>
</tbody>
</table>

* Assuming 3SUM conjecture
CQ*s Domain Split - Progress

Non Acyclic Free-Connex CQ*s*

$O(\log n)$ time $\otimes, \oplus$  

$O(\log n)$ time $\otimes, \oplus$  
Every atom $\phi_i$ contains either all or none of $x_{i+1}, \ldots, x_k$

Order contains a disruptive trio*

* No self-joins, assuming BMM, HYPERCLIQUE conjectures
Local Annotations

• Can we do better with aggregation?
• Annotations derived from aggregation have a specific structure:
  • Only a single relation is annotated with non-identity values
• When only a single relation is annotated with non-identity values, we say the database is **locally-annotated**
Idempotency

• Some semirings have an $\oplus$ operation with a unique property.
• An operation $\oplus$ will be said to be idempotent if for every $x$
  $$x \oplus x = x$$
• Notable commutative semirings with idempotent $\oplus$:

| (Q $\cup \{\infty\}$, $+$, Min, 0, $\infty$) | Min |
| (Q $\cup \{-\infty\}$, $+$, Max, 0, $-\infty$) | Max |
| (Ω, $\cap$, $\cup$, Ω, ∅) | CountD* |

* $O(\log n)$ domain
Idempotent Dichotomy

- \(\text{CQ}^* Q (\vec{x}, \star, \vec{z}) \): \(\neg \phi_1 (\vec{x}, \vec{y}, \vec{z}), \ldots, \phi_l (\vec{x}, \vec{y}, \vec{z})\):
  - Free-connex acyclic
  - No disruptive trio when ignoring \(\star\)
  - Logarithmic-time semiring
  - \(\oplus\)-idempotent semiring

- For CQs we have a known dichotomy

Assuming it is possible to generate an infinite number of domain elements in ascending order
CQ*s Domain Split - Final

Non Acyclic Free-Connex CQ*s*

$O(\log n)$ time $\otimes, \oplus$

$O(\log n)$ time $\otimes, \oplus$

Every atom $\phi_i$ contains either all or none of $x_{i+1}, \ldots, x_k$

Order contains a disruptive trio*

* No self-joins, assuming BMM, HYPERCLIQUE conjectures
ACQs Domain Split - Final

**Non Acyclic Free-Connex ACQs**

- $\alpha$ is one of Min, Max, Sum, Count, Avg
- Every atom $\phi_i$ contains either all or none of $x_{i+1}, \ldots, x_k$

**Count Distinct**
- Sum, Count

Order contains a disruptive trio*

* No self-joins, assuming BMM, HYPERCLIQUE conjectures
Thanks For Listening

Any questions?