Query Evaluation With Constant Delay

Wojciech Kazana

INRIA Saclay, ENS de Cachan

PhD Thesis Defense

LSV, École normale supérieure de Cachan

September 16, 2013
Introduction

Enumeration

Examples

Results

Conclusions
Introduction – databases

Databases: storage of data and retrieval of information.

1. A store has its list of offered products.
   Can I buy orange shoes?

2. Private collection of photos.

3. Map of a metro system.


5. . . .
Introduction – databases

Databases: storage of data and retrieval of information.

1. A store has its list of offered products.
2. Private collection of photos.
   On how many of my photos am I actually present?
3. Map of a metro system.
5. . . .
Introduction – databases

Databases: storage of data and retrieval of information.

1. A store has its list of offered products.
2. Private collection of photos.
3. Map of a metro system. Can I get from Château d’Eau to Bagneux with just one hop?
5. . . .
Introduction – databases

Databases: storage of data and retrieval of information.

1. A store has its list of offered products.
2. Private collection of photos.
3. Map of a metro system.
4. Social network and its graph. Which pairs of people are in the 2-handshakes distance from each other?
5. ...
Query Evaluation Problem:

Input:
- query $q(\overline{x})$
- database $D$ of size $|D|$

Output:
- $q(D)$

Which pairs of people are in the 2-handshakes distance from each other?

$(1, 2), (1, 3), (1, \text{Wojtek}), (1, D), (1, E), (1, P), (1, Q), (1, R)$
$(2, 1), (2, 3), (2, \text{Wojtek}), (2, D), (2, E), (2, P), (2, Q), (2, R)$
$(3, 1), (3, 2), (3, \text{Wojtek}), (3, D), (3, E), (3, P), (3, Q), (3, R) \ldots$
Query Evaluation Problem:

Input:
• query \( q(\bar{x}) \)
• database \( D \) of size \( \|D\| \)

Output:
\[ q(D) \]

Special case: \( q \) boolean = Model Checking Problem.

Are there two green friends? No
Query Evaluation Problem:

Input:
• query $q(\bar{x})$
• database $D$

Output:

$|q(D)| = O(|D|^k)$ if $q$ has $k$ free variables.

$|D|^k$ is too big!
Query Enumeration and Related Problems

Input:
- query $q(\overline{x})$
- database $D$

**Enumeration:**
- compute first solution quickly,
- compute the rest with minimal *delay* between consecutive ones.

**Aim:** First solution in $O(||D||)$, $O(1)$ delay $\rightarrow$ **CONSTANT-DELAY** _lin_
Query Enumeration and Related Problems

Input:
• query $q(\bar{x})$
• database $D$

Enumeration:
• compute first solution quickly,
• compute the rest with minimal delay between consecutive ones.

Aim: First solution in $O(\|D\|)$, $O(1)$ delay $\rightarrow$ CONSTANT-DELAY\textsubscript{lin}

In practice:
- the $O(\|D\|)$ preprocessing is a linear refactorization of the input database (usually adding to it some additional navigational power),
- the refactorized database can then be traversed efficiently, producing new solutions after only constant delays.
Query **Evaluation** Problem:

**Input:**
- query $q(\bar{x})$
- database $D$

**Output:** $q(D)$

Which pairs of people are in the 2-handshakes distance from each other?

(1, 2), (1, 3), (1, Wojtek), (1, D), (1, E), (1, P), (1, Q), (1, R), (2, 1), (2, 3), (2, Wojtek), (2, D), (2, E), (2, P), (2, Q), (2, R), (3, 1), (3, 2), (3, Wojtek), (3, D), (3, E), (3, P), (3, Q), (3, R) ...
Query Enumeration Problem:

Input:
- query $q(\bar{x})$
- database $D$

Output:
$q(D)$

Which pairs of people are in the 2-handshakes distance from each other?

$(1, 2)$
Query Enumeration Problem:

Input:
- query $q(\overline{x})$
- database $D$

Output:
- $q(D)$

Which pairs of people are in the 2-handshakes distance from each other?
- $(1, 2)$, $(1, 3)$
Query Enumeration Problem:

Input:
- query $q(\bar{x})$
- database $D$

Output:

Which pairs of people are in the 2-handshakes distance from each other?

$(1, 2), (1, 3), (1, \text{Wojtek})$
Query Enumeration Problem:

Input:
- query $q(\bar{x})$
- database $D$

Output:

Which pairs of people are in the 2-handshakes distance from each other?

$(1, 2), (1, 3), (1, \text{Wojtek}), (1, D)$
Query Enumeration Problem:

Input:
- query $q(\bar{x})$
- database $D$

Output:

Which pairs of people are in the 2-handshakes distance from each other?

(1, 2), (1, 3), (1, Wojtek), (1, D), (1, E), (1, P), (1, Q), (1, R)
(2, 1), (2, 3), (2, Wojtek), (2, D), (2, E), (2, P), (2, Q), (2, R)
(3, 1), (3, 2), (3, Wojtek), (3, D), (3, E), (3, P), (3, Q), (3, R) ...
Counting Problem:

Input:
- query $q(\bar{x})$
- database $D$

Output: $|q(D)|$

Aim: $O(|D|)$

How many pairs of people are in the 2-handshakes distance from each other? 78
Testing Problem:

Input:
- query $q(\bar{x})$
- database $D$

Dynamical output:
given $\bar{v}$, answer $\bar{v} \in q(D)$

Aim:
- preprocessing (once, $\bar{v}$ unknown) $O(|D|)$
- answering (multiple times) $O(1)$

Is $(1, P)$ in the 2-handshakes distance? Yes
Is $(A, E)$ in the 2-handshakes distance? No
\textbf{Remark 1}

\textbf{CONSTANT-DELAY} \textit{lin} \textit{enumeration} $\rightarrow O(\|D\| + |q(D)|)$ \textit{evaluation}.

\textbf{Remark 2}

\textbf{CONSTANT-DELAY} \textit{lin} \textit{enumeration} $\rightarrow O(\|D\|)$ \textit{model checking}.
Computational model – RAM machine

- Necessary, since we want to talk about linear time.
- We assume that the elements can be compared in constant time.

\[ A <_{\text{lex}} P <_{\text{lex}} \text{Wojtek} \]

In real life: user = (short) e-mail address

- We can sort lexicographically tuples of constant size in linear time.

Radix sort

- We can follow pointers in constant time.

Direct access to the \( n \)-th cell of an array.

- Coding of a graph:

List of consecutive edges.
NOT an adjacency matrix!
Introduction

Enumeration

Examples

Results

Conclusions
Example 1: enumeration of edges

Database:

\[ \text{graph } G = (V, E) \]
\[ \|G\| = |V| + |E| \]

- **Constant-Delay** enumeration is not too difficult.

- \(O(\|G\|)\) counting is not too difficult.

- **Testing** requires logarithmic time.
  - \(O(1)\) testing if \(G\) is a tree.

Query:

\[ q(x, y) = E(x, y) \]
Example 2: complement of a graph

Database:

graph $G = (V, E)$

$\|G\| = |V| + |E|$

Query:

$q(x, y) = \neg E(x, y)$

- **Constant-Delay** enumeration already not trivial.

- $O(\|G\|)$ counting is not too difficult.

- Testing still requires logarithmic time.
  - $O(1)$ testing if $G$ is a tree.
Example 2: complement of a graph

Is given by the following list:

\[(1, 1)\] \(\rightarrow\) \[(2, 1)\] \(\rightarrow\) \[(3, 2)\] \(\rightarrow\) \[(3, 4)\]

\[(1, 4)\] \(\rightarrow\) \[(3, 1)\] \(\rightarrow\) \[(3, 3)\] \(\rightarrow\) \[(4, 4)\]
Example 2: complement of a graph

(1,1) → (1,2) → (1,3) → (1,4)

(2,1) → (2,2) → (2,3) → (2,4)

(3,1) → (3,2) → (3,3) → (3,4)

(4,1) → (4,2) → (4,3) → (4,4)
Example 2: complement of a graph
Example 3: 2-handshake distance

Database:

\[ \text{graph } G = (V, E) \]
\[ \|G\| = |V| + |E| \]

Query:

\[ q(x, y) = \exists z E(x, z) \land E(z, y) \]

- **CONSTANT-DELAY\_lin enumeration** not possible? (Bagan’09)
  - **CONSTANT-DELAY\_lin** enumeration if \( G \) has bounded degree.

- **\( O(\|G\|) \)** counting not possible?
  - **\( O(\|G\|) \)** counting if \( G \) has bounded degree.

- **\( O(1) \)** testing not possible? (Bagan’09)
  - **\( O(1) \)** testing if \( G \) has bounded degree.
Introduction

Enumeration

Examples

Results

Conclusions
Classes of graphs

- Bounded tree-width
- Planar
- Excluded minor
- Excluded topological minor
- Bounded degree
- Bounded expansion
(K, Segoufin’11) gives a new proof of:

**Theorem 1 (Durand, Grandjean’07)**

\[ C \text{ with bounded degree, } q(\bar{x}) \in \text{FO}, \text{ given } G \in C, \]

the *enumeration* of \( q \) over \( G \) is in \text{CONSTANT-DELAY} \text{lin}.

Moreover, the output is returned in the lexicographical order.

*The hidden constants are triply exponential in } |q| \ (K, Segoufin’11).
Results - FO over structures with bounded degree

(K, Segoufin’11) gives new proofs of:

**Theorem 2 (Bagan, Durand, Grandjean, Olive’08)**

\[ C \text{ with bounded degree}, \ q(\bar{x}) \in \text{FO}, \text{ given } G \in C, \text{ the counting} |q(D)| \text{ is in } O(|G|). \]

**Theorem 3 (Lindell’08)**

\[ C \text{ with bounded degree}, \ q(\bar{x}) \in \text{FO}, \text{ given } G \in C, \text{ the testing of } q \text{ over } G \text{ is in } O(1) \text{ after } O(|G|) \text{ preprocessing}. \]

The hidden constants are triply exponential in $|q|$ (K, Segoufin’11).
Comments

• The original proof:
  ◦ Quantifier elimination procedure based on bijective representation.
  ◦ Non-elementary dependency on $|q|$ ($2^{\frac{2}{|q|}}$).

• Our proof:
  ◦ Based on Gaifman locality of FO.
  ◦ Gives the $2^{2^{|q|}}$ dependency on $|q|$. 
Results - FO over structures with bounded expansion

Theorem 4 (K, Segoufin’13)

\[ \mathcal{C} \text{ with bounded expansion, } q(\bar{x}) \in \text{FO}, \text{ given } G \in \mathcal{C}, \]

the enumeration of \( q \) over \( G \) is in \( \text{CONSTANT-DELAY}_{\text{lin}} \).

Moreover, the output is returned in the lexicographical order.
Theorem 5 (K, Segoufin’13)

\( \mathcal{C} \) with bounded expansion, \( q(\bar{x}) \in \text{FO} \), given \( G \in \mathcal{C} \),
the \textbf{counting} \(|q(D)|\) is in \( O(\|G\|) \).

Theorem 6 (K, Segoufin’13)

\( \mathcal{C} \) with bounded expansion, \( q(\bar{x}) \in \text{FO} \), given \( G \in \mathcal{C} \),
the \textbf{testing} of \( q \) over \( G \) is in \( O(1) \) after \( O(\|G\|) \) preprocessing.
Comments

- The hidden constants are non-elementary \((2^{2^\cdots^2})\). 
  \[ |q| \]

- This is unavoidable already for model checking over unranked trees, unless \(\text{FPT} = \text{AW}[\ast]\) (Frick, Grohe’04).
Proof Strategy

- Quantifier elimination procedure:
  - For all $q(\bar{x}y)$ quantifier free,
  - Exists $q'(\bar{x})$ quantifier free s.t.
  - Given $G \in C$, in $O(|G|)$ we construct $G'$ s.t.
  - $(\exists y q)(G) = q'(G')$.

Graph $G'$ is an “augmentation” of $G$, which allows us to continue the inductive process.

- We then solve the quantifier free case.

Both steps are not trivial.
(K, Segoufin’12) gives a new proof of:

**Theorem 7 (Bagan’06)**

\[ C \text{ with bounded treewidth, } q(\bar{x}) \in \text{MSO}, \text{ given } G \in C, \]
the **enumeration** of \( q \) over \( G \) is in \( \text{CONSTANT-DELAY}_{\text{lin}} \).
Results - MSO over structures with bounded treewidth

(K, Segoufin’12) gives new proofs of:

Theorem 8 (Arnborg, Lagergren, Seese’91)

\[ C \text{ with bounded treewidth, } q(\bar{x}) \in \text{MSO}, \text{ given } G \in C, \text{ the counting } |q(D)| \text{ is in } O(\|G\|). \]

Theorem 9

\[ C \text{ with bounded treewidth, } q(\bar{x}) \in \text{MSO}, \text{ given } G \in C, \text{ the testing of } q \text{ over } G \text{ is in } O(1) \text{ after } O(\|G\|) \text{ preprocessing}. \]
Comments

- The original proof:
  - Allows for monadic second-order free variables.
  - No bound on total memory usage during the enumeration phase.
  - Rather complicated reasoning concerning tree automatons.

- Our proof:
  - Only first-order free variables.
  - Constant total memory usage during the enumeration phase.
  - Sequence of reduction steps.

- In both cases the hidden constants are non-elementary \( (2^{2^{|q|}}) \).
MSO enumeration – outline

Theorem 6 (Bagan’06)

$C$ with bounded treewidth, $q(\bar{x}) \in \text{MSO}$, given $G \in C$, the \textit{enumeration} of $q$ over $G$ is in \textsc{Constant-Delay}_{\text{lin}}.

The proof of (K, Segoufin’12) is a sequence of consecutive reduction steps:
MSO enumeration – outline

Reduction steps:

1. Trees instead of structures of bounded tree-width.
   Compute the tree decomposition in linear time. (Bodlaender)
   Interpret the tree decomposition in MSO. (Courcelle)
2. Binary trees instead of unranked trees.
   First child – next sibling encoding. (Rabin)
3. Ancestor-typed outputs including all the least common ancestors.
   Composition Lemma.
5. Binary queries from $\Sigma_2(<)$.
   Colcombet.
Step 3: Ancestor-typed outputs including all the least common ancestors.

$q(x,y)$

$q'(x,y,z)$

$q'(x,y,z)$
Step 4: Binary queries.

\[ q(x, y, z) = \bigvee_{q', q''} q'(x, y) \land q''(x, z). \]

Disjunction is exclusive.

Composition Lemma for MSO over trees (can be proved using a simple Ehrenfeucht-Fraïssé game argument).
Step 5: Binary queries from $\Sigma_2(\prec)$.

$s'(z, x)$ and $s''(z, y)$ are of the form $\exists \bar{v} \forall \bar{u} \theta(x, y, z, \bar{v}, \bar{u})$, where $\theta$ is quantifier free.

Theorem 10 (Colcombet)

Over binary trees, every MSO formula $q(x, y)$ is equivalent to a $\Sigma^+_2(\prec)$ formula $q'(x, y)$. $q' = \exists \bar{v} \forall \bar{u} \theta(x, y, z, \bar{v}, \bar{u})$, where $\theta$ is a disjunction of conjunctions of atomic predicates or MSO queries with one free variable or atoms using $\prec$. 
The rest is an induction on the number of free variables:

\[ q(x, y, z, u) = q'(x, y, z) \land q''(x, u) \]

- We inductively enumerate \( q'(x, y, z) \).
- For every solution \((a, b, c)\) to \( q' \) we (efficiently) extend it with all solutions to \( q'' \) of the form \((a, \_\)\).
Summary

Bounded tree-width
MSO - CDlin, N-Elem

Planar

Bounded degree
FO - CDlin, 3-EXP

Excluded minor

Excluded topological minor

Bounded expansion
FO - CDlin, N-Elem

Nowhere dense

Somewhere dense

FO - MC not in FPT (Dawar, Kreutzer)
Push further along the lines of the current approach:

- Consider FO queries over classes of nowhere dense structures.
- Consider other query languages over structures with particular properties.
**Perspectives**

\textbf{Constant-Delay}_{\textit{lin}} implies evaluation algorithm working in time \(O(\|G\| + |q(G)|).\)

- When is the converse true?

When is \textbf{Constant-Delay}$_{\textit{lin}}$ enumeration impossible?

- Most of the lower bounds require complexity assumptions (matrix multiplication, FPT \(\neq AW[\ast],\) etc.).
- In the mentioned cases also \(O(\|G\| + |q(G)|)\) evaluation is not possible.
- How much can we prove directly?
Perspectives

Let $A, B \in \text{CONSTANT-DELAY}_{lin}$ (black boxes).

Under what assumptions

- $A \cup B \in \text{CONSTANT-DELAY}_{lin}$, *(The best understood.)*
- $A \cap B \in \text{CONSTANT-DELAY}_{lin}$,
- $\bar{A} \in \text{CONSTANT-DELAY}_{lin}$?
Thank You!