Constant delay enumeration for FO queries over databases with local bounded expansion

Luc Segoufin [1] **Alexandre Vigny**[2]

[1]ENS Ulm, Paris

[2]Université Paris Diderot, Paris

October 16, 2017

Introduction

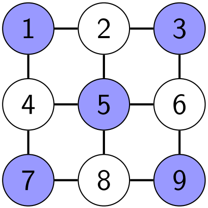
- Query q *small*
- Database D *huge*
- Compute $q(D)$ *gigantic*

Examples :

query q

$$q(x,y) := \exists z(B(x) \wedge E(x,z) \wedge \neg E(y,z))$$

database D



solutions $q(D)$

$\{(1,2)\ (1,3)\ (1,4)$
$(1,6)\ (1,7)\ \cdots$
$(3,1)\ (3,2)\ (3,4)$
$(3,6)\ (3,7)\ \cdots$
$\cdots\}$

Enumeration

Input : $\|D\| := n$ & $\|q\| := k$ (computation with RAM)

Goal : output solutions one by one (no repetition)

- STEP 1 : Preprocessing

  Prepare the enumeration : Database $D \longrightarrow$ Index I

  Preprocessing time : $f(k) \cdot n \rightsquigarrow O(n)$

- STEP 2 : Enumeration

  Enumerate the solutions : Index $I \longrightarrow \overline{x_1}, \overline{x_2}, \overline{x_3}, \overline{x_4}, \cdots$

  Delay : $O(f(k)) \rightsquigarrow O(1)$

Constant delay enumeration after linear preprocessing

Example 1

Input :

- Database $D := \langle\{1, \cdots, n\}; E\rangle$ $\qquad \|D\| = |E| \quad (E \subseteq D \times D)$
- Query $q(x, y) := \neg E(x, y)$

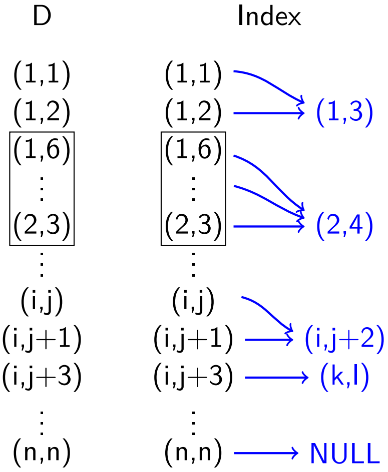
D

(1,1)
(1,2)
(1,6)
⋮
(2,3)
⋮
(i,j)
(i,j+1)
(i,j+3)
⋮
(n,n)

Example 1

Input :

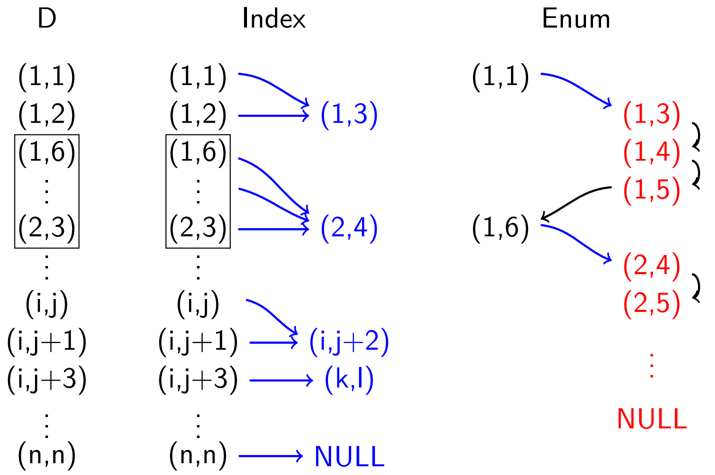
- Database $D := \langle\{1, \cdots, n\}; E\rangle$ $\quad\quad \|D\| = |E| \quad (E \subseteq D \times D)$
- Query $q(x,y) := \neg E(x,y)$



Example 1

Input :

- Database $D := \langle\{1,\cdots,n\}; E\rangle$ $\quad \|D\| = |E| \quad (E \subseteq D \times D)$
- Query $q(x,y) := \neg E(x,y)$



Example 2

Input :

- Database $D := \langle \{1, \cdots, n\}; E_1; E_2 \rangle \quad \|D\| = |E_1| + |E_2| \quad (E_i \subseteq D \times D)$
- Query $q(x,y) := \exists z,\ E_1(x,z) \wedge E_2(z,y)$

Example 2

Input :

- Database $D := \langle \{1, \cdots, n\}; E_1; E_2 \rangle \quad \|D\| = |E_1| + |E_2| \quad (E_i \subseteq D \times D)$
- Query $q(x,y) := \exists z,\ E_1(x,z) \wedge E_2(z,y)$

B : Adjacency matrix of E_2

$$\begin{pmatrix} E_2(1,1) & \dots & E_2(1,y) & \dots & E_2(1,n) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ E_2(z,1) & \dots & E_2(z,y) & \dots & E_2(z,n) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ E_2(n,1) & \dots & E_2(n,y) & \dots & E_2(n,n) \end{pmatrix}$$

$$\begin{pmatrix} E_1(1,1) & \dots & E_1(1,i) & \dots & E_1(1,n) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ E_1(x,1) & \dots & E_1(x,z) & \dots & E_1(x,n) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ E_1(n,1) & \dots & E_1(n,z) & \dots & E_1(n,n) \end{pmatrix} \quad \begin{pmatrix} q(1,1) & \dots & q(1,y) & \dots & q(1,n) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ q(x,1) & \dots & q(x,y) & \dots & q(x,n) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ q(n,1) & \dots & q(n,y) & \dots & q(n,n) \end{pmatrix}$$

A : Adjacency matrix of E_1

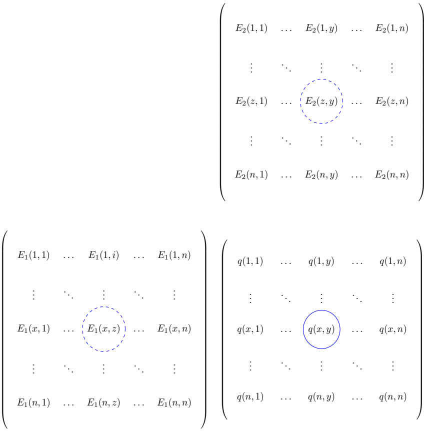
C : Result matrix

Example 2

Input :

- Database $D := \langle\{1,\cdots,n\}; E_1; E_2\rangle$ $\quad \|D\| = |E_1| + |E_2| \quad (E_i \subseteq D \times D)$
- Query $q(x,y) := \exists z,\ E_1(x,z) \land E_2(z,y)$

B : Adjacency matrix of E_2



A : Adjacency matrix of E_1 $\qquad$ C : Result matrix

Compute the set of solutions

=

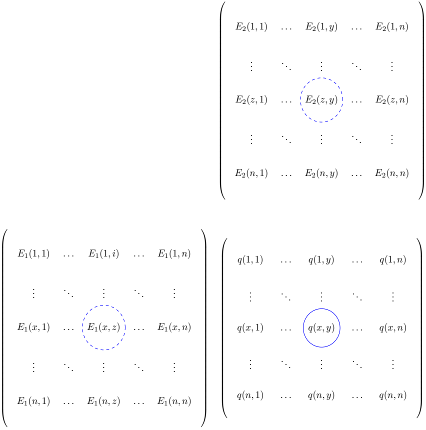
boolean matrix multiplication

Example 2

Input :

- Database $D := \langle\{1, \cdots, n\}; E_1; E_2\rangle$ $\quad \|D\| = |E_1| + |E_2| \quad (E_i \subseteq D \times D)$
- Query $q(x, y) := \exists z,\ E_1(x, z) \wedge E_2(z, y)$

B : Adjacency matrix of E_2



A : Adjacency matrix of E_1

C : Result matrix

- Linear preprocessing : $O(n^2)$
- Number of solutions : $O(n^2)$
- Algorithm for the boolean matrix multiplication in $O(n^2)$
- Conjecture :
  "There are no algorithm for the boolean matrix multiplication working in time $O(n^2)$."

Example 2

Input :

- Database $D := \langle \{1, \cdots, n\}; E_1; E_2 \rangle \quad \|D\| = |E_1| + |E_2| \quad (E_i \subseteq D \times D)$
- Query $q(x,y) := \exists z,\ E_1(x,z) \wedge E_2(z,y)$

This query cannot be enumerated with constant delay [1]

1. Unless there is a breakthrough with the boolean matrix multiplication.

Example 2

Input :

- Database $D := \langle\{1, \cdots, n\}; E_1; E_2\rangle \quad \|D\| = |E_1| + |E_2| \quad (E_i \subseteq D \times D)$
- Query $q(x, y) := \exists z,\ E_1(x, z) \wedge E_2(z, y)$

This query cannot be enumerated with constant delay [1]

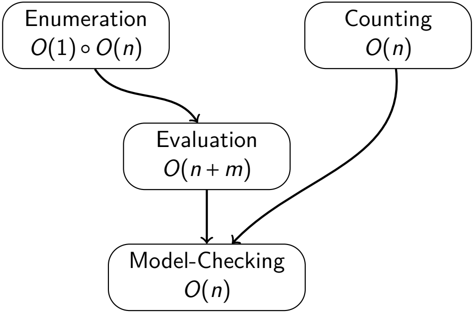
We need to put restrictions on queries and/or databases.

1. Unless there is a breakthrough with the boolean matrix multiplication.

Other problems

For FO queries over a class $\mathscr{C}$ of databases.

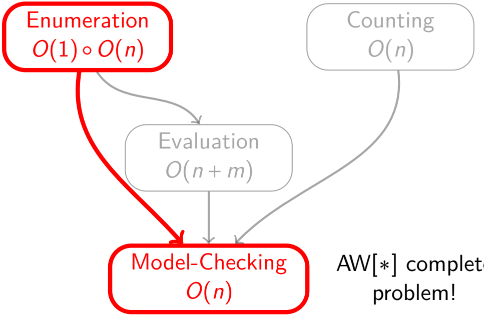
Model-Checking	:	Is this true ?	$O(n)$
Enumeration	:	Enumerate the solutions	$O(1) \circ O(n)$
Counting	:	How many solutions ?	$O(n)$
Evaluation	:	Compute the entire set	$O(n+m)$



Other problems

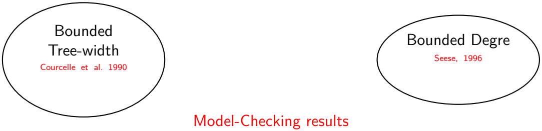
For FO queries over a class $\mathscr{C}$ of databases.

Model-Checking	:	Is this true ?	$O(n)$
Enumeration	:	Enumerate the solutions	$O(1) \circ O(n)$
Counting	:	How many solutions ?	$O(n)$
Evaluation	:	Compute the entire set	$O(n+m)$

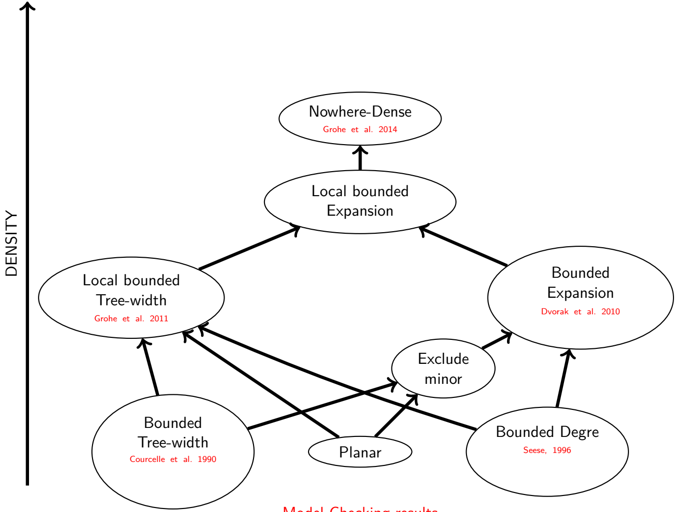


AW[∗] complete problem!

Classes of graphs closed under taking sub-graphs

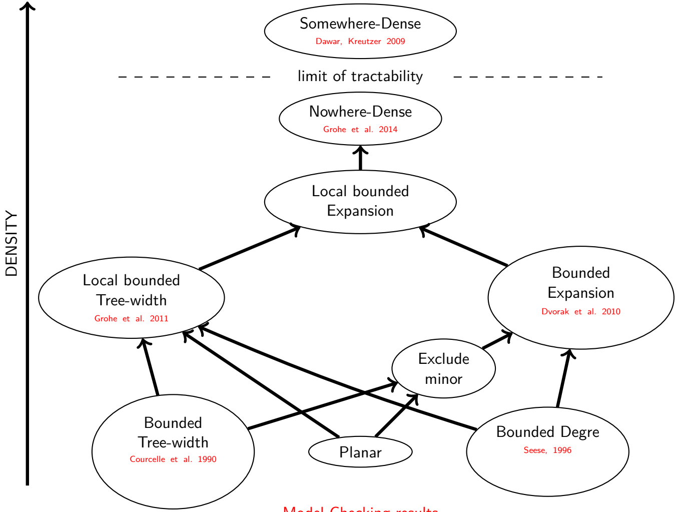


Classes of graphs closed under taking sub-graphs



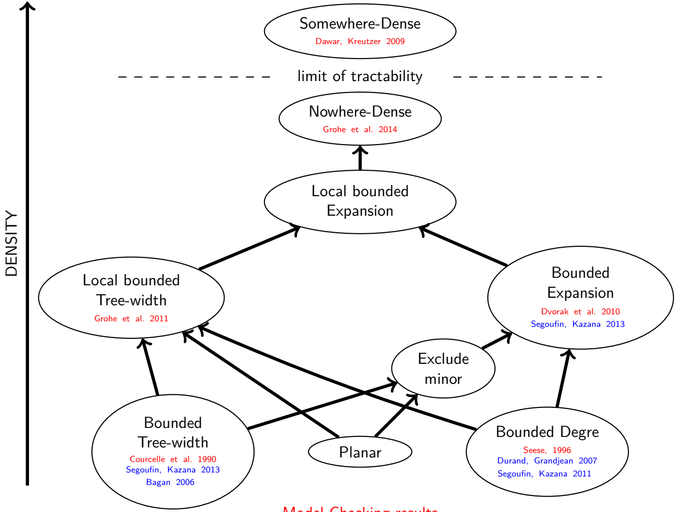
Model-Checking results

Classes of graphs closed under taking sub-graphs



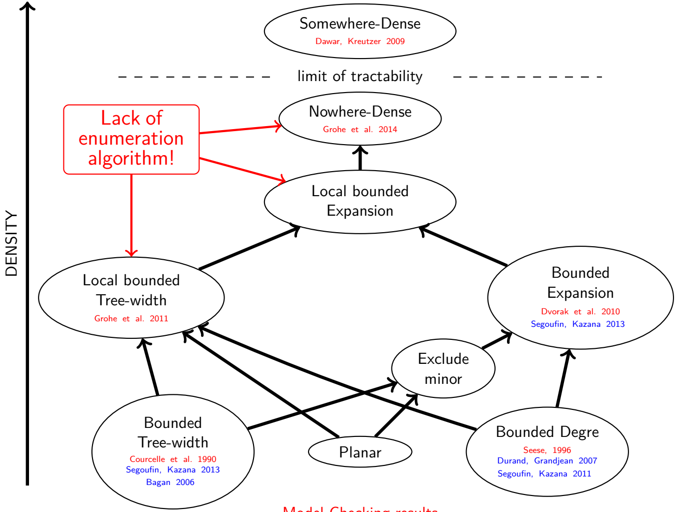
Model-Checking results

Classes of graphs closed under taking sub-graphs



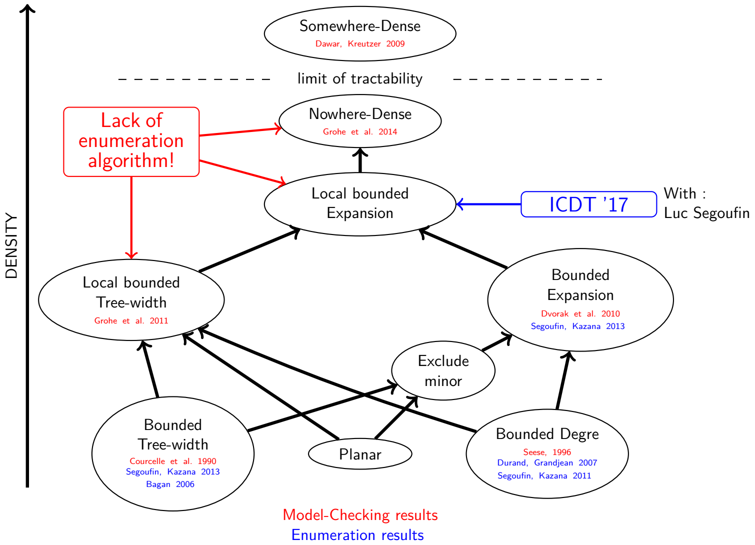
Model-Checking results
Enumeration results

Classes of graphs closed under taking sub-graphs



Model-Checking results
Enumeration results

Classes of graphs closed under taking sub-graphs



Local bounded expansion ?

Definition : Class of r-neighborhoods

Let $\mathscr{C}$ be a class of graphs, $r \in \mathbb{N}$, $\mathscr{C}_r := \{N_r^G(a) \mid G \in \mathscr{C},\ a \in G\}$

Definition : Local bounded expansion

$\mathscr{C}$ has locally bounded expansion if for all r, $\mathscr{C}_r$ as bounded expansion.

Local bounded expansion ?

Definition : Class of r-neighborhoods

Let $\mathscr{C}$ be a class of graphs, $r \in \mathbb{N}$, $\mathscr{C}_r := \{N_r^G(a) \mid G \in \mathscr{C},\ a \in G\}$

Definition : Local bounded expansion

$\mathscr{C}$ has locally bounded expansion if for all r, $\mathscr{C}_r$ as bounded expansion.

Bounded expansion ?

Local bounded expansion ?

Definition : Class of r-neighborhoods

Let $\mathscr{C}$ be a class of graphs, $r \in \mathbb{N}$, $\mathscr{C}_r := \{N_r^G(a) \mid G \in \mathscr{C},\ a \in G\}$

Definition : Local bounded expansion

$\mathscr{C}$ has locally bounded expansion if for all r, $\mathscr{C}_r$ as bounded expansion.

Bounded expansion ?

Examples

Planar graphs, graphs with bounded degree, bounded tree width, ...

Properties

Bounded in-degree, linear number of edges, nice coloring, ...

Local bounded expansion ?

Definition : Class of r-neighborhoods

Let $\mathscr{C}$ be a class of graphs, $r \in \mathbb{N}$, $\mathscr{C}_r := \{N_r^G(a) \mid G \in \mathscr{C},\ a \in G\}$

Definition : Local bounded expansion

$\mathscr{C}$ has locally bounded expansion if for all r, $\mathscr{C}_r$ as bounded expansion.

Can have a non linear number of edges !

Bounded expansion ?

Examples

Planar graphs, graphs with bounded degree, bounded tree width, ...

Properties

Bounded in-degree, linear number of edges, nice coloring, ...

Our results

Theorem (Segoufin, V. 17')

Over classes of graphs with *local bounded expansion*, for every FO query, after a pseudo-linear preprocessing, we can :

- enumerate with constant delay every solutions.
- test in constant time whether a given tuple is a solution.
- compute in constant time the number of solutions.

Pseudo-linear ?

A function f is pseudo linear if and only if :

$$\forall \epsilon > 0, \quad \exists N_\epsilon \in \mathbb{N}, \quad \forall n \in \mathbb{N}, \quad n > N_\epsilon \implies f(n) \leq n^{1+\epsilon}$$

$$n \ll n\log^i(n) \ll \text{pseudo-linear} \ll n^{1,0001} \ll n\sqrt{n}$$

"Pseudo-linear $\approx n\log^i(n)$"

"Pseudo-constant $\approx \log^i(n)$"

Tools used

We use :

- Gaifman normal form for FO queries.
- Neighbourhood cover.[1]
- Enumeration for graphs with **Bounded expansion**.[2]
- New short-cut pointers dedicated to the enumeration.

1. Grohe, Kreutzer, Siebertz '14
2. Segoufin, Kazana. '13

Future/Current work

- The nowhere-dense case !
- Enumeration with update :
  What happens if a small change occurs after the preprocessing ?

  Existing results for : words, graphs with bounded tree-width or bounded degree.

Future/Current work

- The nowhere-dense case !
- Enumeration with update :
  What happens if a small change occurs after the preprocessing ?

  Existing results for : words, graphs with bounded tree-width or bounded degree.

Thank you !

Questions ?