Enumerating Answers to Ontology-Mediated Queries

Carsten Lutz

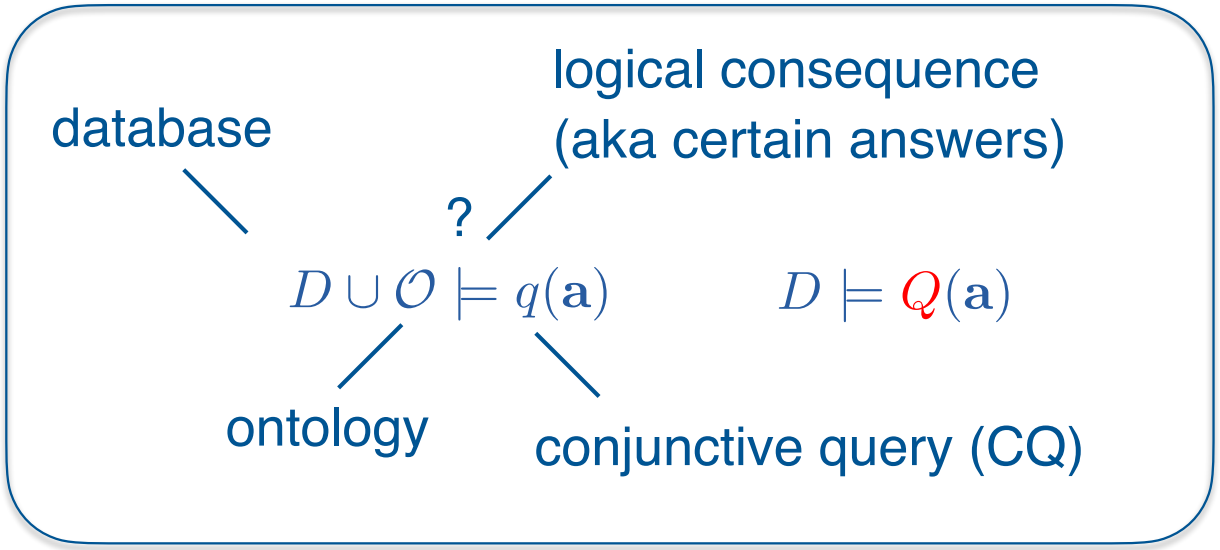
Universität Leipzig

Joint work with Marcin Przybylko (PODS22 + a bit of AAAI23)

Ontology-Mediated Queries

Ontology is logical theory, formalizes and provides domain knowledge

Ontology-mediated querying:



CQ for instance: $q(x) = \exists y(\ \mathsf{ComputerScientist}(x) \wedge \mathsf{collaboratedWith}(x, y) \wedge \mathsf{Biologist}(y)\ )$

Ontology-mediated query (OMQ) $Q(\bar{x}) = (\mathcal{O}, \Sigma, q)$ consists of

ontology $\mathcal{O}$, data schema Σ, actual query q

Ontologies and TGDs

Tuple-generating dependencies (TGDs) take form

$$\forall \bar{x} \forall \bar{y} \, (\, \phi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \, \psi(\bar{x}, \bar{z}) \,)$$

Here we work with guarded TGDs:

$$R(x, y, z) \wedge S(x, y) \wedge A(z) \rightarrow \exists \bar{u} \, \psi(x, y, \bar{u})$$

Ontologies are sets of guarded TGDs (generalizes $\mathcal{ELI}$)

$$\mathsf{Movie}(x) \rightarrow \exists y \exists z \, \mathsf{directedBy}(x, y) \wedge \mathsf{Director}(y) \wedge \mathsf{hasLocation}(x, z) \wedge \mathsf{GeoLocation}(z)$$

$$\mathsf{Movie}(x) \wedge \mathsf{hasScene}(x, y) \wedge \mathsf{Violent}(y) \rightarrow \exists z \, \mathsf{hasRating}(x, z) \wedge \mathsf{AgeRestriction}(z)$$

Enumerating OMQ Answers

Main problem studied:

Enumerate all answers to OMQ Q on database D, without repetition.

Focus on enumeration in CD∘Lin:

- preprocessing phase takes time linear in $|D|$
- enumeration delay between two answers independent of $|D|$ ('constant')

This refers to data complexity: $|Q|$ considered a constant

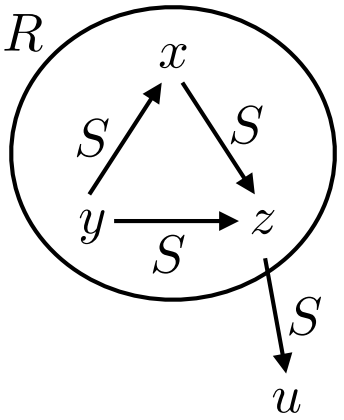
Notes:

- first answer produced after time $f(|Q|) \cdot O(|D|)$,
  thus fixed-parameter linear (FPL) to decide whether answer exists
- set of all answers produced in time linear in $|D| + |Q(D)|$

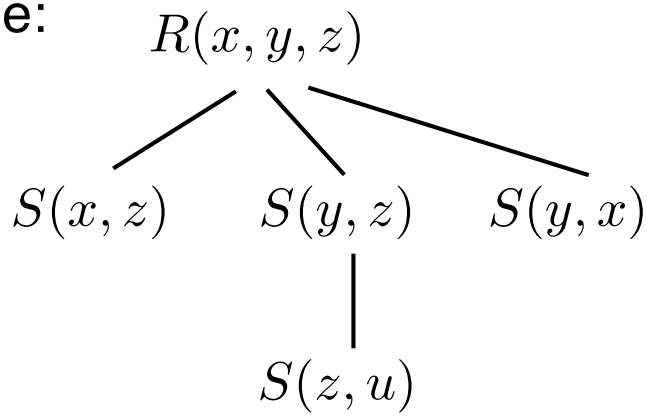
Acyclicity and Free-connex Acyclicity

CQ is acyclic if it has a join tree

CQ $q(x, y, z)$:



join tree:



Same as generalized hypertree width 1

CQ $q(\bar{x})$ is free-connex acyclic if $q(\bar{x}) \wedge R(\bar{x})$ is acyclic

Acyclicity and Free-connex Acyclicity

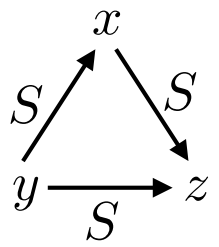
Acyclicity and free-connex acyclicity are independent notions

Acyclic, but not free-connex acyclic:

$$q(x, y) = x \xrightarrow[S]{} z \xrightarrow[S]{} y$$

Free-connex acyclic, but not acyclic:

$$q(x, y, z) =$$



Will use these notions also for OMQs, mean query component

Enumeration in CD∘Lin

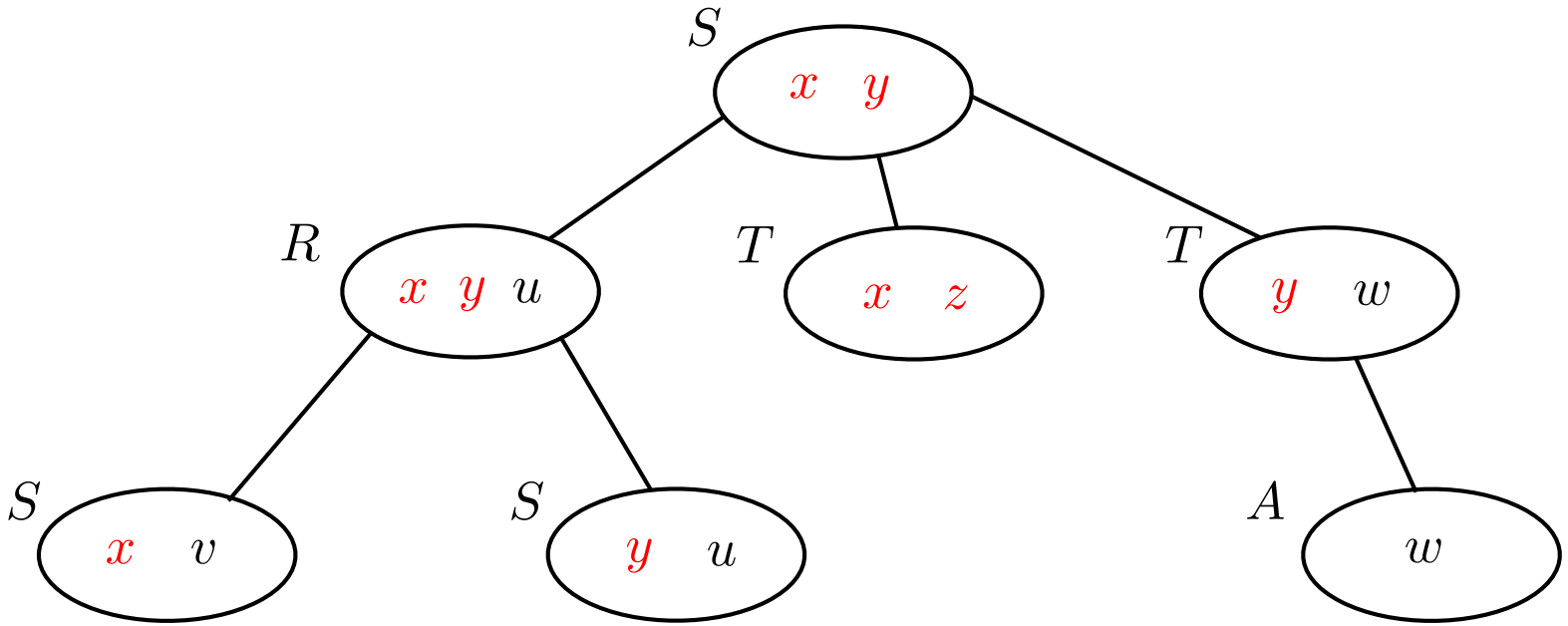
Theorem [BaganDurandGrandjean2007]

For CQs that are acyclic and free-connex acyclic, enumeration is in CD∘Lin.

Also see excellent enumeration tutorial [BerkholzGerhardtSchweikardt20]

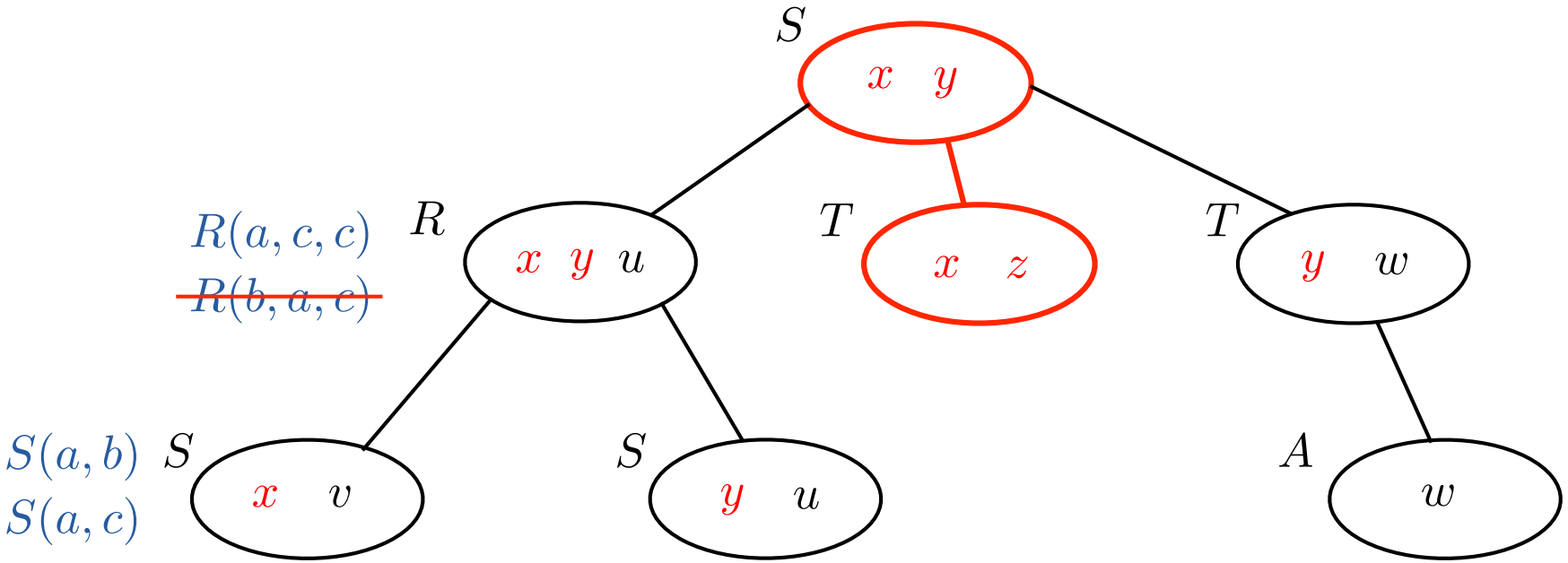
Enumeration in CD°Lin

Compute (generalized) join tree, prefix contains exactly answer variables:



Enumeration in CD°Lin

Compute (generalized) join tree, prefix contains exactly answer variables:

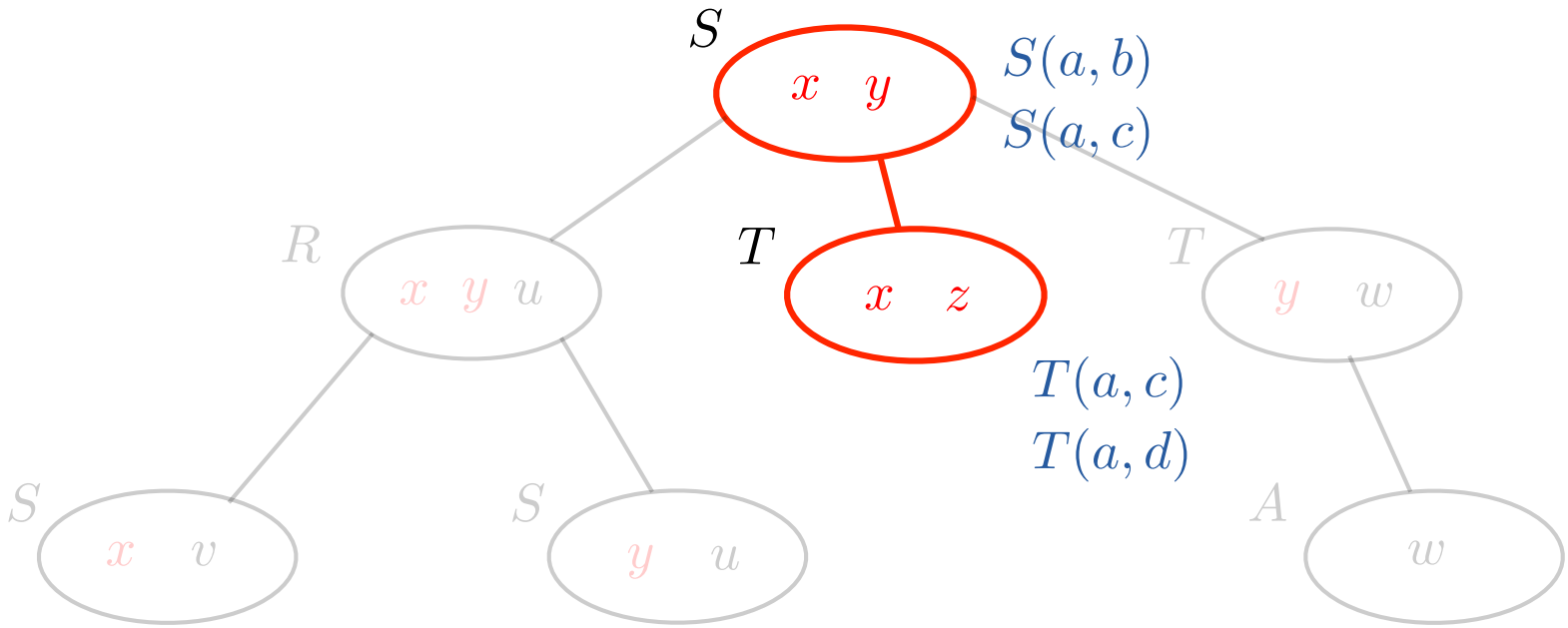


Materialize relevant relations in the nodes (single scan through *D*)

Bottom-up pass builds semi-join between parents and children

Enumeration in CD°Lin

Compute (generalized) join tree, prefix contains exactly answer variables:



Materialize relevant relations in the nodes (scan though *D*)

Bottom-up pass builds semi-join between parents and children

Pre-order tree walk assembles answers, output at final node, backtrack

Constant delay relies on pre-computed indexes during preprocessing phase

With Ontologies

Bagan et al.'s result lifts to OMQs based on guarded TGDs:

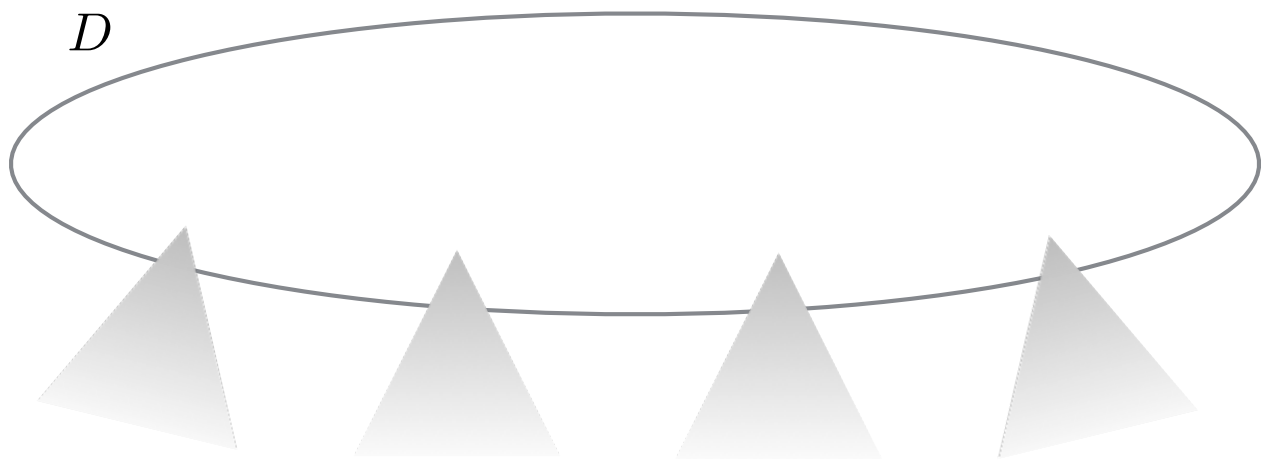
Theorem

For OMQs from $(\mathrm{GTGD}, \mathrm{CQ})$ that are acyclic and free-connex acyclic, enumeration is in $\mathrm{CD}\circ\mathrm{Lin}$.

This is a significant generalization, also includes recursive queries

Reducing Out Ontologies

The chase: start from database, apply TGDs from ontology as rules

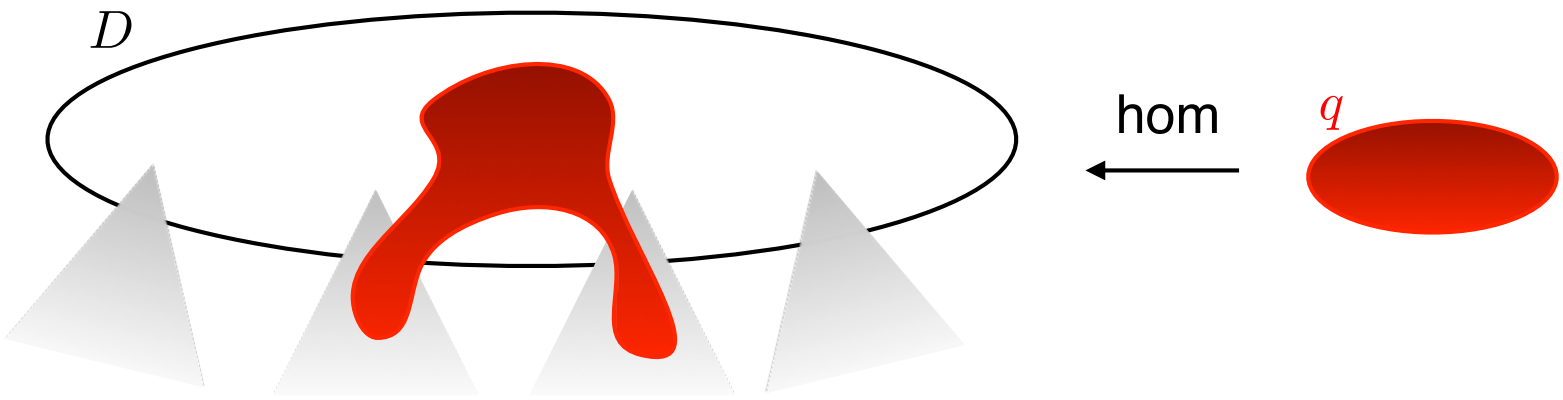


Chased database

- gives correct answers (we can forget ontology)
- may be infinite
- has a very regular shape

Reducing Out Ontologies

Only certain finite parts of chase matter:



Any homomorphism from q to $\mathsf{ch}_{\mathcal{O}}(D)$ gives rise to

"excursions" of q into chase trees

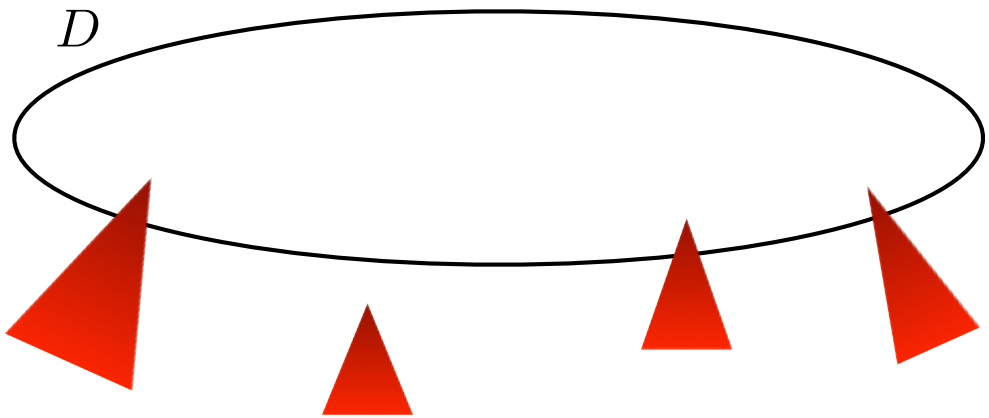
Consider set $\mathsf{cl}(q)$ of tree-like CQs:

start from q, identify variables, take subquery

These CQs, viewed as chase (sub)trees, are all we need to know about chase

Reducing Out Ontologies

We obtain a partial chase:



This chase fragment can be computed in linear time:

- describe database part of chase by propositional Horn formula θ
- compute minimal model for θ in linear time
- construct partial chase from model in straightforward way

We can then disregard $\mathcal{O}$ and enumerate answers to q
on partial chase using blackbox procedure

Testing

All Testing:

Preprocessing given Q and D, then get + test answer candidates $\bar{c}_1, \bar{c}_2, \ldots$

CD∘Lin defined in obvious way

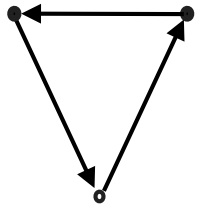
Same result as for enumeration, but with different condition:

- enumeration: acyclic and free-connex acyclic
- all-testing: free-connex acyclic

Also considered single testing in linear time

here CQ needs to be weakly acyclic

= acyclic after replacing answer vars with constants



Partial Answers

New notion of answer to OMQ:

- answer can have form $(a, *, b, *, *, a)$
- wildcard $*$ denotes constant whose exact identity is unknown

Example: $\mathsf{Researcher}(x) \rightarrow \exists y\, \mathsf{hasOffice}(x, y)$

$\mathsf{Researcher}(\mathsf{mary})$

$q(x, y) = \mathsf{Researcher}(x) \wedge \mathsf{hasOffice}(x, y)$

No complete answers, but partial answer $(\mathsf{mary}, *)$

Minimally partial answer (MPA):

partial answer $(a, *, b, *, *, a)$ and no strictly more informative answer

such as $(a, *, b, c, *, a)$

Enumerating MPAs

Theorem

For OMQs from (GTGD, CQ) that are acyclic and free-connex acyclic, enumerating MPAs is in $\text{CD}\circ\text{Lin}$.

Naive approaches will not work:

- in pre-order tree walk: replace answer variables also with wildcards

  fails: joins need exact identities of constants

- in pre-order tree walk: replace answer variables also with nulls

  fails: will produce duplicate answers

Enumerating MPAs

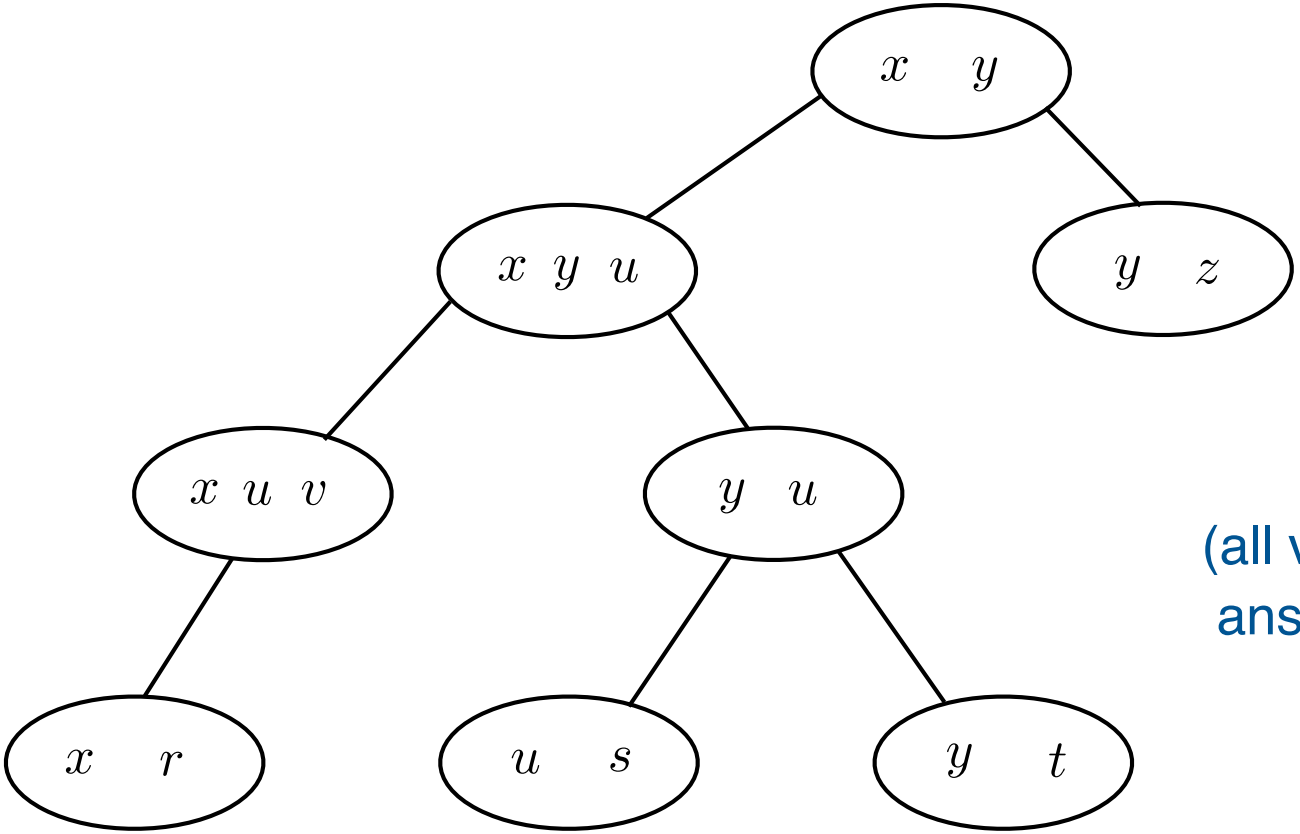
Theorem

For OMQs from (GTGD, CQ) that are acyclic and free-connex acyclic, enumerating MPAs is in CD$\circ$Lin.

Central ideas of enumeration algorithm:

- Preprocessing phase: construct partial chase of database,
  then execute all preprocessing steps as described before
  (materialize, build semi-joins, remove quantified variables)
- Preprocessing phase:
  precompute „possible excursions“ of query into existential part of chase
  and arrange resulting excursion trees in suitable data structure
- Enumeration phase:
  use excursion trees to produce wildcard parts of answers
  prune for minimality

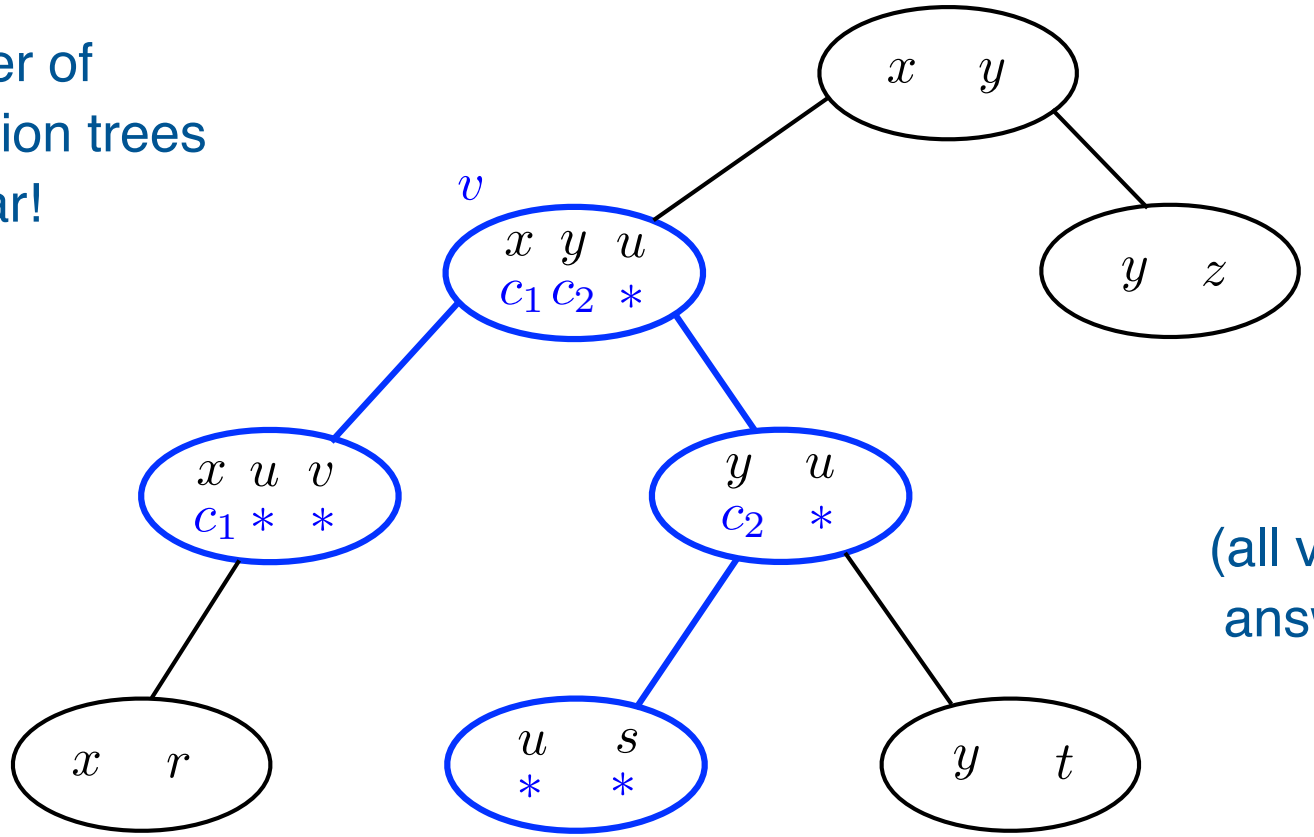
Excursion Trees



(all variables
answer variables)

Excursion Trees

Number of excursion trees is linear!



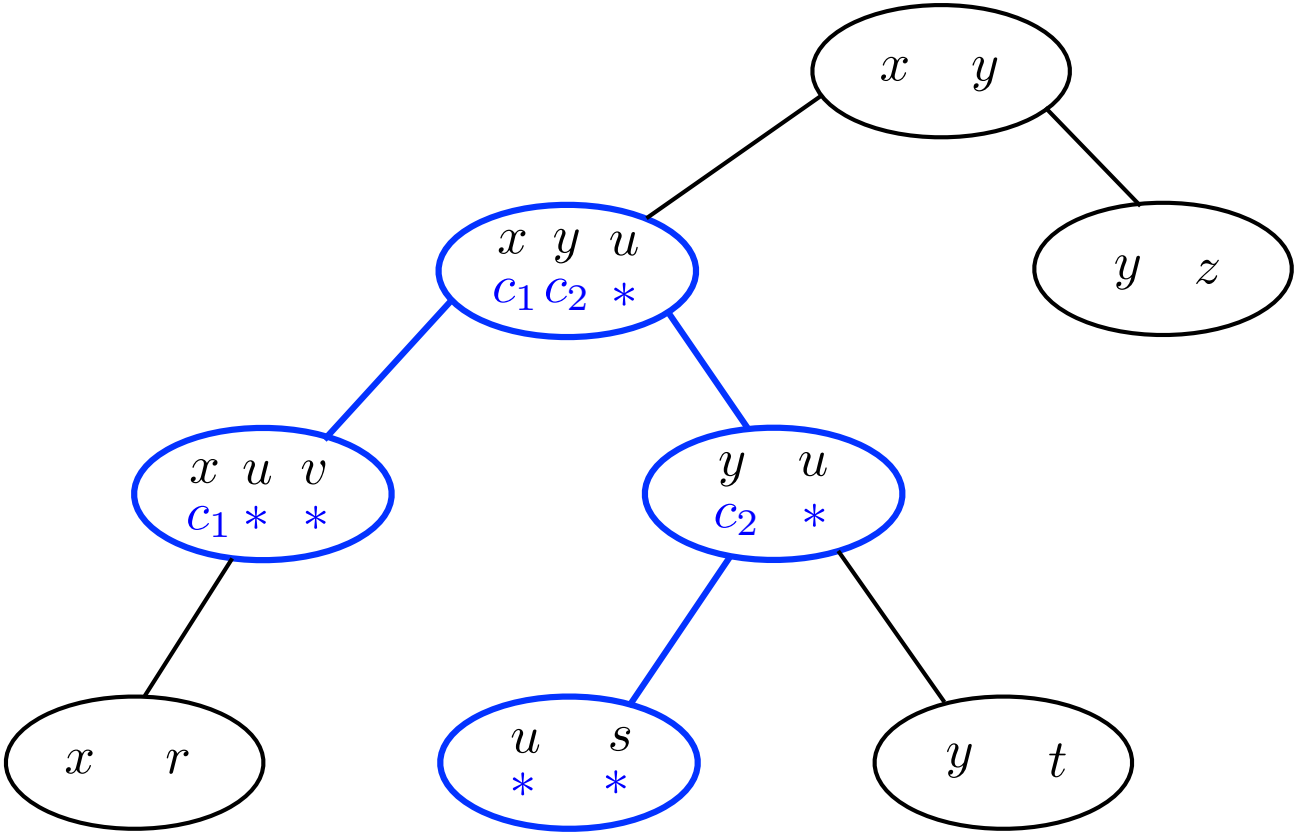
(all variables answer variables)

Preprocessing: Compute lists $\mathsf{ptree}(v, h)$ of all excursion trees that

- are rooted at v and
- map 'predecessor variables' in v according to h

Sort them in `database-preferring order' (i.e.: disfavour $*$)

Excursion Trees

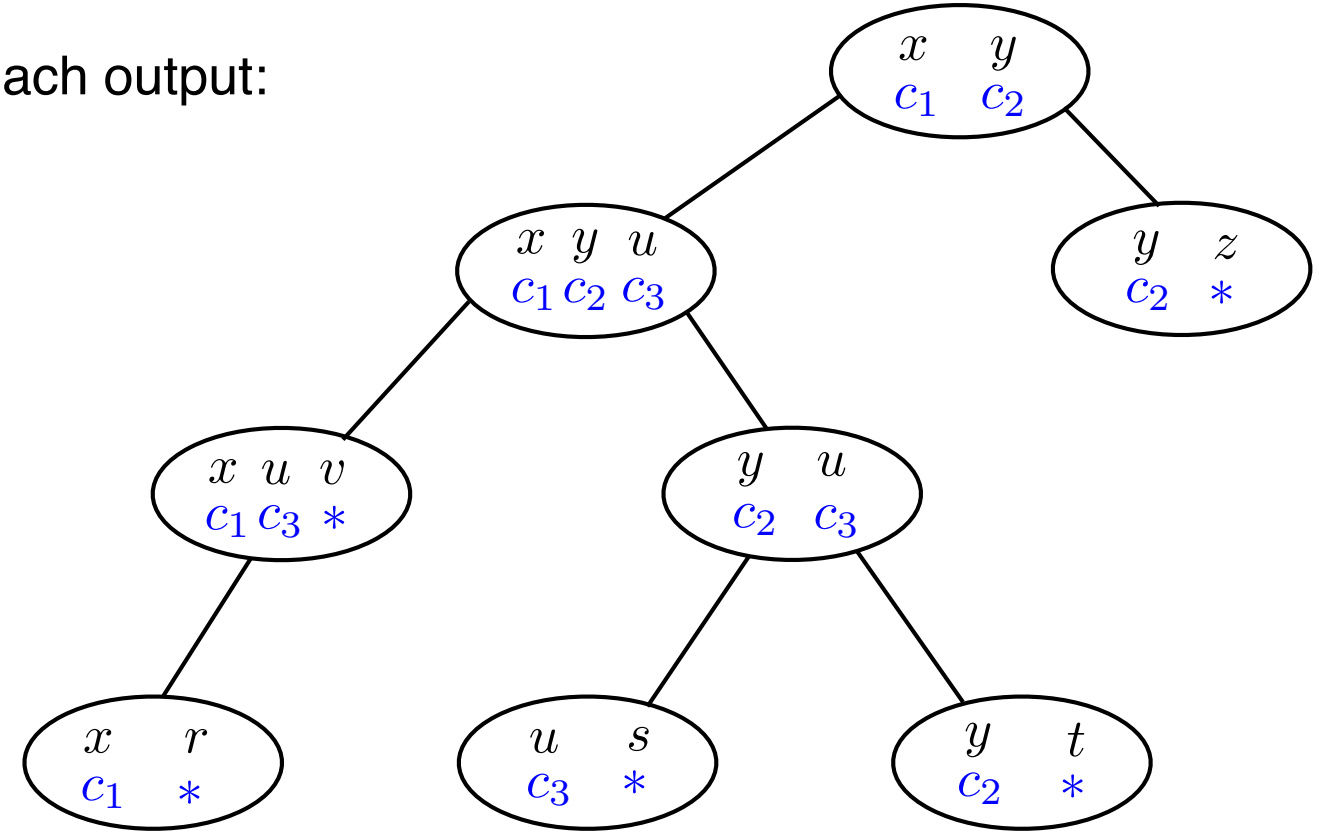


Enumeration phase:

Excursion trees induce jumps in pre-order tree walk

Pruning

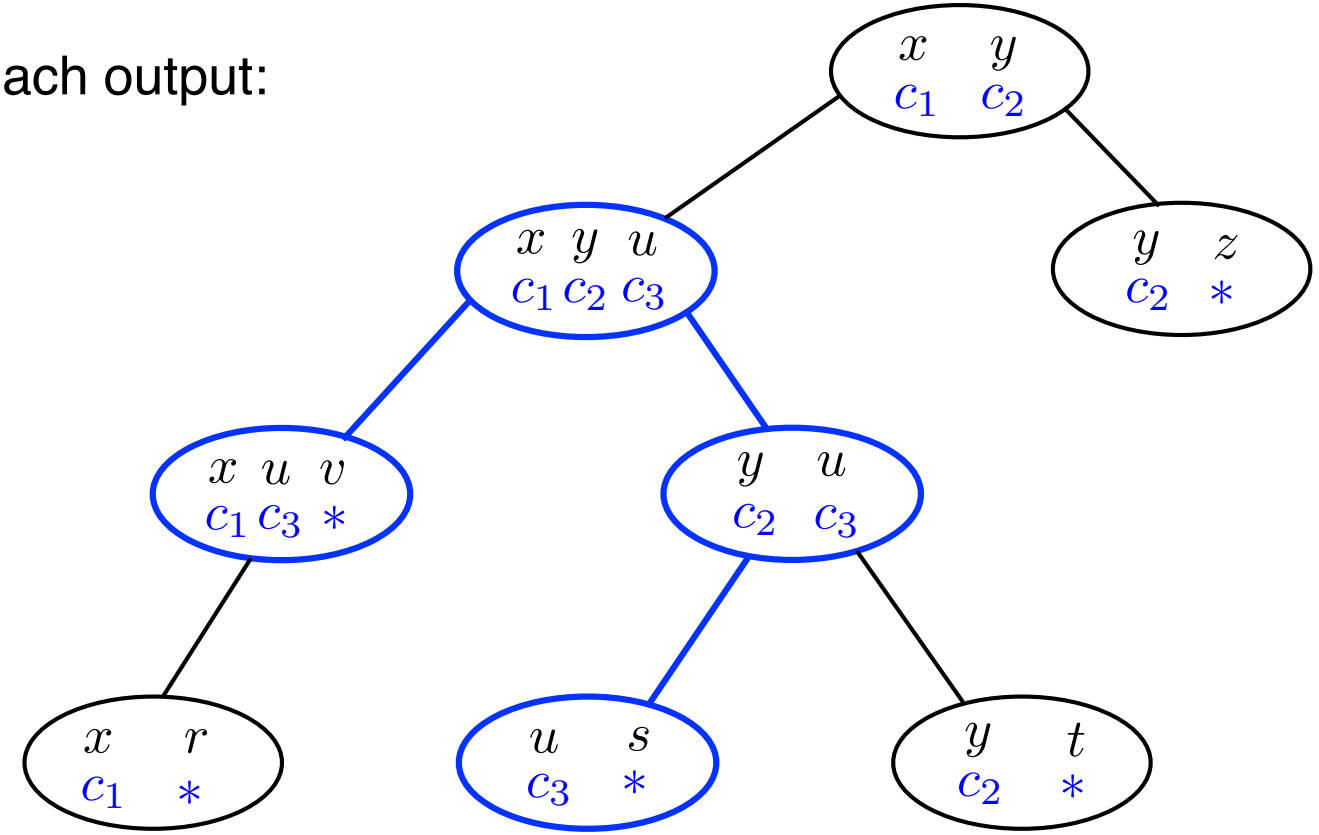
After each output:



Consider all subtrees (not necessarily excursion trees)

Pruning

After each output:

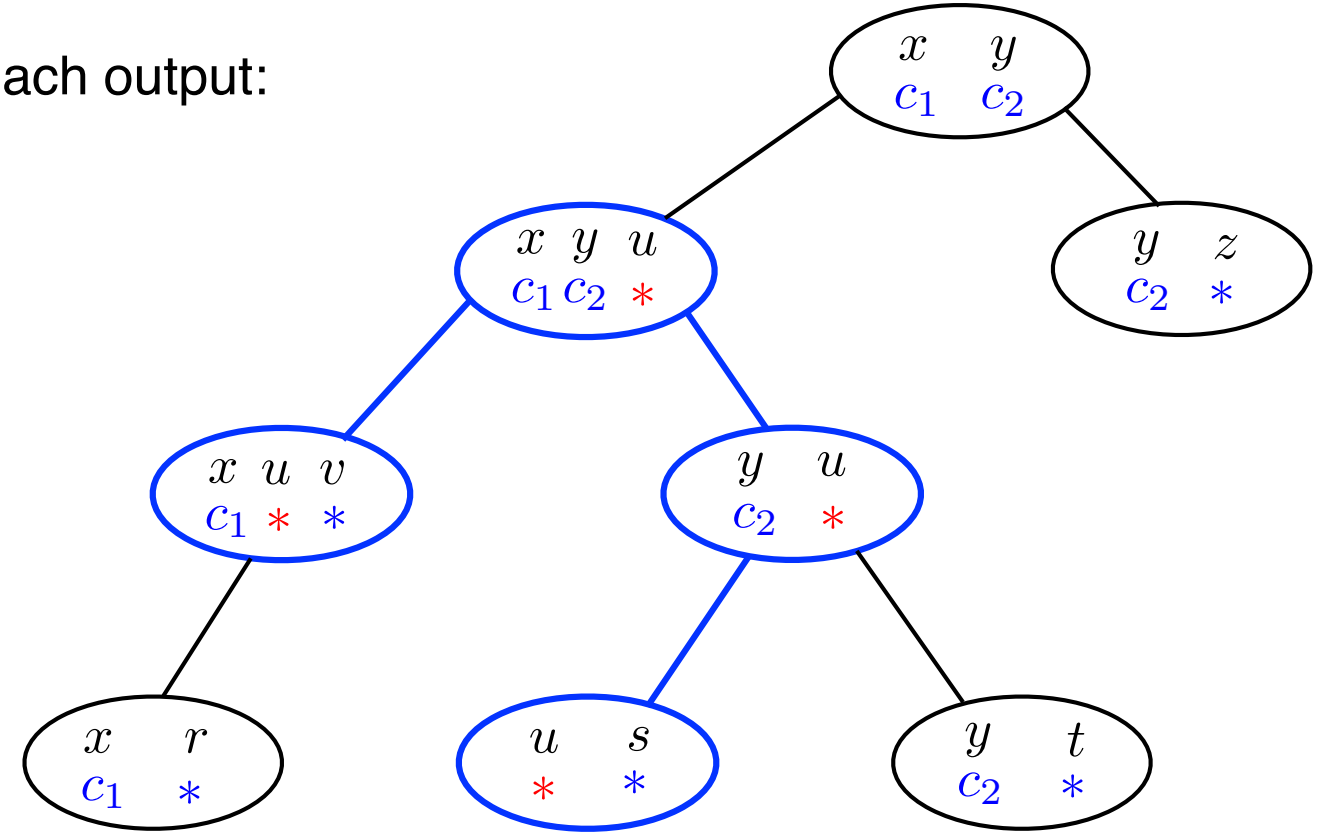


Consider all subtrees (not necessarily excursion trees)

and all ways of replacing constants with $*$

Pruning

After each output:



Consider all subtrees (not necessarily progress trees)

and all ways of replacing constants with $*$

Remove all excusion trees obtained in this way from lists $\mathsf{ptree}(v, h)$

Minimality

Only minimally partial answers are output because:

- excursion trees are sorted in database-preferring order

  $\Rightarrow$ less partial answers output first

- pruning removes dominated excursion trees

  $\Rightarrow$ non-minimally partial answers never produced

Multiple Wildcards

Stronger notion of MPAs to OMQs:

- answer can have form $(a, *_1, b, *_2, *_1, a)$
- intuition: adds equality between wildcards (but not inequality)

Theorem

For OMQs from (GTGD, CQ) that are acyclic and free-connex acyclic, enumeration of MPAs with multi-wildcards is in CD$\circ$Lin.

Challenge: Wildcards may be shared among different excursion trees

Enumeration by non-trivial combination of

- enumeration of minimally partial answers with single wildcard
- all-testing of ~~minimally~~ partial answers with multi-wildcards

  (ah, well, a slight variation thereof)

Multiple Wildcards

First idea:

- use algorithm for MPAs with single wildcard as blackbox
- when tuple $\bar{a}$ is output
  - replace '$*$' with '$*_i$', $i \geq 1$, in all possible ways
  - use all-testing to filter out tuples that are not partial answers
  - output minimally partial tuples from remaining set

Problem:

- assume that (a, b) is an answer
- then $(*, *)$ is not an MPA
- but $(*_1, *_1)$ might be, is incomparable to (a, b)!

Multiple Wildcards

First idea:

- use algorithm for MPAs with single wildcard as blackbox
- when tuple $\bar{a}$ is output
  - replace '$*$' with '$*_i$', $i \geq 1$, in all possible ways
  - use all-testing to filter out tuples that are not partial answers
  - output minimally partial tuples from remaining set

Multiple Wildcards

Hint to solution:

- use algorithm for MPAs with single wildcard as blackbox
- when tuple $\bar{a}$ is output
  - replace '$*$' with '$*_i$', $i \geq 1$, in all possible ways
  - use all-testing to filter out tuples that are not partial answers
  - output minimally partial tuples from remaining set

Multiple Wildcards

Hint to solution:

- use algorithm for MPAs with single wildcard as blackbox
- when tuple $\bar{a}$ is output
  - first (possibly) replace constants in $\bar{a}$ with wildcard '$*$'

    replace '$*$' with '$*_i$', $i \geq 1$, in all possible ways

    for instance: if $\bar{a} = (a, b)$, we produce also $(*, *)$, then $(*_1, *_1)$
  - use all-testing to filter out tuples that are not partial answers
  - output minimally partial tuples from remaining set

Reality is more subtle, have to make sure that
no duplicates are generated and there is always something to ouput

Lower Bounds

Theorem [Bagan et al 2007, Brault-Baron 2013]

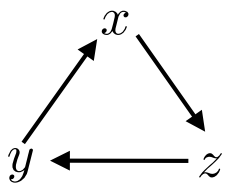
Let q be a CQ that is self-join free.

1. If q is not acyclic, then enumerating answers to q is not in CD∘Lin unless triangle conjecture (or generalization to hypercliques) fails
2. If q is acyclic but not free-connex acyclic, then enumerating answers to q is not in CD∘Lin unless Sparse Boolean Matrix Multiplication is in $O(|M_1| + |M_2| + |M_1 M_2|)$

Triangle conjecture:
triangle detection in undirected graphs not possible in linear time

From fine-grained complexity theory [AbboudVassilevskaWilliams14]

E.g. cyclic query $q(x, y, z) =$



Triangle detection possible after first answer

Lower Bounds

Theorem [Bagan et al 2007, Brault-Baron 2013]

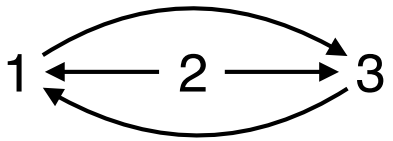
Let q be a CQ that is self-join free.

1. If q is not acyclic, then enumerating answers to q is not in CD∘Lin unless triangle conjecture (or generalization to hypercliques) fails
2. If q is acyclic but not free-connex acyclic, then enumerating answers to q is not in CD∘Lin unless Sparse Boolean Matrix Multiplication is in $O(|M_1| + |M_2| + |M_1 M_2|)$

Paradigmatic query to encode BMM: $q(x, y) = \exists z\, x \xrightarrow[r_1]{} z \xrightarrow[r_2]{} y$

representation of matrix M_i:

	1	2	3
1	0	0	1
2	1	0	1
3	1	0	0



$M_1 M_2(i, j) = 1$ if there is k such that $M_1(i, k) = M_2(k, j) = 1$

With Ontologies

Theorem

Let Q be an OMQ from $(\mathcal{ELI}, \text{CQ})$ that is non-empty and self-join free.

1. If Q is not acyclic, then enumerating answers to Q is not in CD∘Lin unless (a generalization of) the triangle conjecture fails.
2. If Q is acyclic and connected, but not free-connex acyclic, then enumerating answers to Q is not in CD∘Lin unless Sparse Boolean Matrix Multiplication is in $O(|M_1| + |M_2| + |M_1 M_2|)$

$\mathcal{ELI}$ cannot be replaced with GTGD:
then we could remove 'self-join free' using ontology statements $R'(\bar{x}) \leftrightarrow R(\bar{x})$

Connected cannot be dropped from Point 2 (have counterexample).

Applies to complete answers and minimal partial answers (both types)

All-Testing MPAs

All-testing is less well behaved for MPAs:

Theorem

There is an OMQ $Q \in (\mathcal{ELI}, \text{CQ})$ that is acyclic and free-connex acyclic s.t. all-testing MPAs to Q is not in CD$\circ$Lin unless the triangle conjecture fails.

Intuitive reason: Testing single answer with wildcards positively may rule out large number of complete answers!

Applies to MPAs with single wildcard and with multi-wildcards

All-Testing MPAs

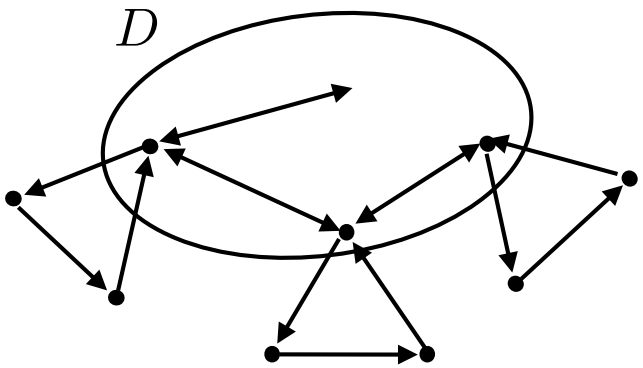
Theorem

There is an OMQ $Q \in (\text{GTGD}, \text{CQ})$ that is acyclic and free-connex acyclic s.t. all-testing MPAs to Q is not in CD$\circ$Lin unless the triangle conjecture fails.

Reduce triangle detection to all-testing MPAs in (GTGD,CQ)

Undirected graph viewed as database with symmetric edges

Ontology $\mathcal{O}$: $E(x,y) \to \exists y_1 \exists y_2\, E(x_1,y_1) \wedge E(y_1,y_2) \wedge E(y_2,x)$



Query $q(x,y,z,u)$:

$x \longrightarrow y \longrightarrow z \longrightarrow u$

for all $c \in \text{dom}(D)$:

$(c,*,*,c)$ is partial answer

Is it minimally partial?

Functional Roles

Functional Roles are important feature of description logic (DL)

For example $\mathcal{ELIHF}$:

Concepts formed according to rule

$$C, D ::= \top \mid A \mid C \sqcap D \mid \exists r.C \mid \exists r^-.C$$

Ontologies are sets of

- concept inclusions $C \sqsubseteq D$
- role inclusions $r \sqsubseteq s$
- functionality assertions $\mathsf{func}(r)$

} GTGDs

~~GTGDs~~

CQs under functionality assertions alone

= CQs under unary functional dependencies [CarmeliKröll2020]

Functional Roles

Theorem [CarmeliKröll2018]

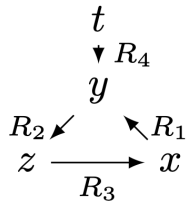
For every CQ q and set of unary FDs Γ such that q^+_Γ is acyclic and free-connex acyclic, enumeration of q under Γ is in CD$\circ$Lin.

q^+ obtained from q by extending atoms with additional variables,

guided by FDs in Γ; for example:

$$q(x,t) = R_1(x,y) \wedge R_2(y,z) \wedge R_3(z,x) \wedge R_4(t,y)$$

$$\Gamma = \{\mathsf{func}(R_2^-), \mathsf{func}(R_3^-), \mathsf{func}(R_4)\}$$



Results in

$$q^+(x,t,y,z) = R_1'(x,y,z), R_2'(y,z), R_3'(z,x,y), R_4'(t,y)$$

q is neither acyclic nor free-connex acyclic, q^+ is both

Functional Roles

Theorem

For OMQs $(\mathcal{O}, q)$ from $(\mathcal{ELIHF}, \text{CQ})$ such that $q^+_{\mathcal{O}}$ is is acyclic and free-connex acyclic, enumeration is in CD$\circ$Lin.

Upper bound:

use partial chase to remove ontology, then transition to q^+,

then use CarmeliKröll as a blackbox

Lower bound:

then use CarmeliKröll as a blackbox

Questions?