Extremal Fitting Problems for Conjunctive Queries

Carsten Lutz
Universität Leipzig

Joint work with
Balder ten Cate, Victor Dalmau, Maurice Funk (mostly PODS23)

CQs and the Fitting Problem

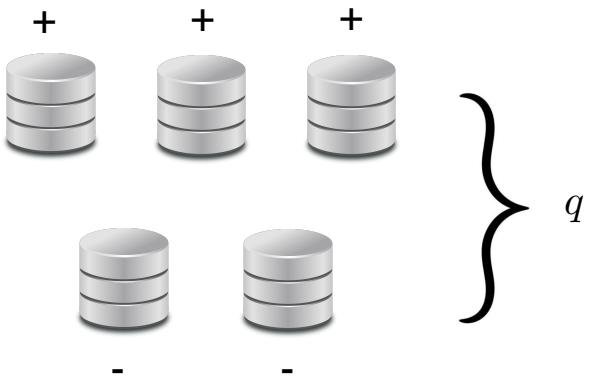
Conjunctive query (CQ) is FO formula using only connectives $\wedge, \exists$, e.g.

$$q(\mathsf{student}, \mathsf{lecturer}) = \exists \mathsf{course}\, \big(\mathsf{attends}(\mathsf{student}, \mathsf{course}) \wedge \mathsf{teaches}(\mathsf{lecturer}, \mathsf{course})\big)$$

Free variables referred to as answer variables

Labeled data example: triple $(D, \bar{a}, \ell)$ with

- D a database instance (finite relational structure)
- $\bar{a}$ a tuple over the active domain / universe of D
- $\ell \in \{+, -\}$ is the label (positive or negative example)



The fitting problem:

Input: collection E of labeled examples

Question: is there a CQ q that fits all examples, i.e. for all $(D, \bar{a}, \ell) \in E$: $\bar{a} \in q(D)$ iff $\ell = +$

Of course one may also ask to compute a concrete such q (if existent)

Motivation

Old school: query by example (QBE)

a user wants to write a query, but is not able to formalize it

they give positive and negative examples and want the query to be derived automatically

New school: machine learning

fundamental theorem of machine learning theory tightly links fitting algorithms to PAC learning

if there is a PAC learner at all, then every algorithm that produces a fitting object is PAC

(but is there a PAC learner for CQs? — we'll see)

Homomorphisms Everywhere

CQ $q(\bar{x})$ can be viewed as finite relational structure with distinguished elements $(q, \bar{x})$

CQ evaluation is about homomorphisms:

for CQ $q(\bar{x})$ and database D: $\bar{a} \in q(D)$ iff $(q, \bar{x}) \longrightarrow (D, \bar{a})$

homomorphism existence

CQ containment is about homomorphisms:

CQ q_1 is contained in CQ q_2, written $q_1 \subseteq q_2$,

if for all databases D: $q_1(D) \subseteq q_2(D)$

Theorem [ChandraMerlin1977]

For all CQs $q_1(\bar{x})$, $q_2(\bar{x})$: $q_1 \subseteq q_2$ iff $(q_2, \bar{x}) \longrightarrow (q_1, \bar{x})$

Fitting Existence and Construction - Example

Consider databases over schema that contains single binary relation

I.e. digraphs Let C_i be the cycle of length i

Try to find a fitting Boolean (i.e. 0-ary) CQ:

$$(C_3,+) \qquad (C_5,+) \qquad (C_7,+) \qquad (C_{105},-)$$

Such a q does not exist:

- q must contain a cycle, otherwise $q \to C_{105}$
- consider any cycle in q, say of length i
- since $q \to C_3$, i is a divisible by 3 …and 5 …and 7
- 3, 5, 7 are prime, so i is also divisible by $3 \cdot 5 \cdot 7 = 105$
- this can be used to show $q \to C_{105}$ ↯ Fittings are related to … products!

The (Basic) Fitting Problem

Theorem [Willard2010 / tenCateDalmau2015]

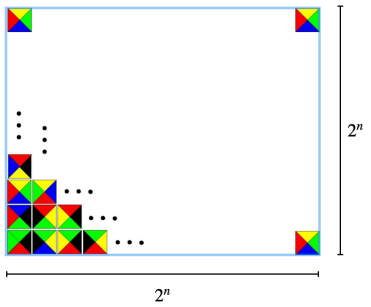
The fitting problem for conjunctive queries is coNExpTime-complete.

Algorithm for upper bound:

- compute direct product of positive examples $(P, \bar{b}) := \prod_{(D,\bar{a},+)\in E} (D, \bar{a})$
- check whether $(P, \bar{b}) \longrightarrow (D, \bar{a})$ for any $(D, \bar{a}, -) \in E$
- return 'no' if this is the case, otherwise $(P, \bar{b})$ viewed as CQ fits all examples

Lower bound by reduction from the $2^n \times 2^n$ tiling problem

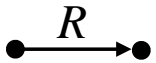
Hints:
- bit-wise decomposition of the $2^n \times 2^n$-grid into n positive examples
- single negative example ensures existence of tiling



Extremal Fitting CQs

In general, there may be many different CQs that fit given set of examples, e.g.

Positive example 

Negative example 

What fitting (Boolean) CQs can you think of?

Do extremal fittings exist? Are they unique?

Are there only finitely many?

How to compute them?

How to verify them?

We can compare fittings q_1, q_2 by query containment:

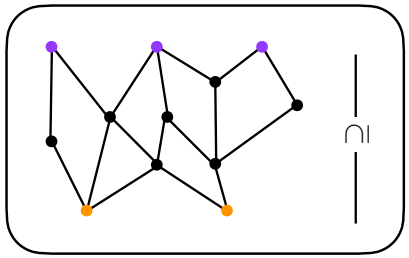
$q_1 \subseteq q_2$ means "q_2 is more general than q_1"

and "q_1 is more specific than q_2"

Two extremes:

- most-general fitting queries
  (no strictly more general query fits)
- most-specific fitting queries
  (no strictly more specific query fits)

Together describe space of all fittings:



In machine learning, this is called version space representation

The Homomorphism Lattice

Homomorphism order $\longrightarrow$ induces (pre)-lattice on set of all CQs (Boolean, fixed schema)

- greatest lower bound of $q_1, \ldots, q_n$: direct product $q_1 \times \cdots \times q_n$
- least upper bound of $q_1, \ldots, q_n$: disjoint union $q_1 \uplus \cdots \uplus q_n$

The structure of this lattice is interesting:

- large parts of it are dense: if $q_1 \longrightarrow q_2$ then we find q with $q_1 \longrightarrow q \longrightarrow q_2$
- the density gaps have been exactly characterized: [NesetrilTardif2000]

  1. every acyclic CQ q_2 gives rise to density gap: q_2 has a covered element

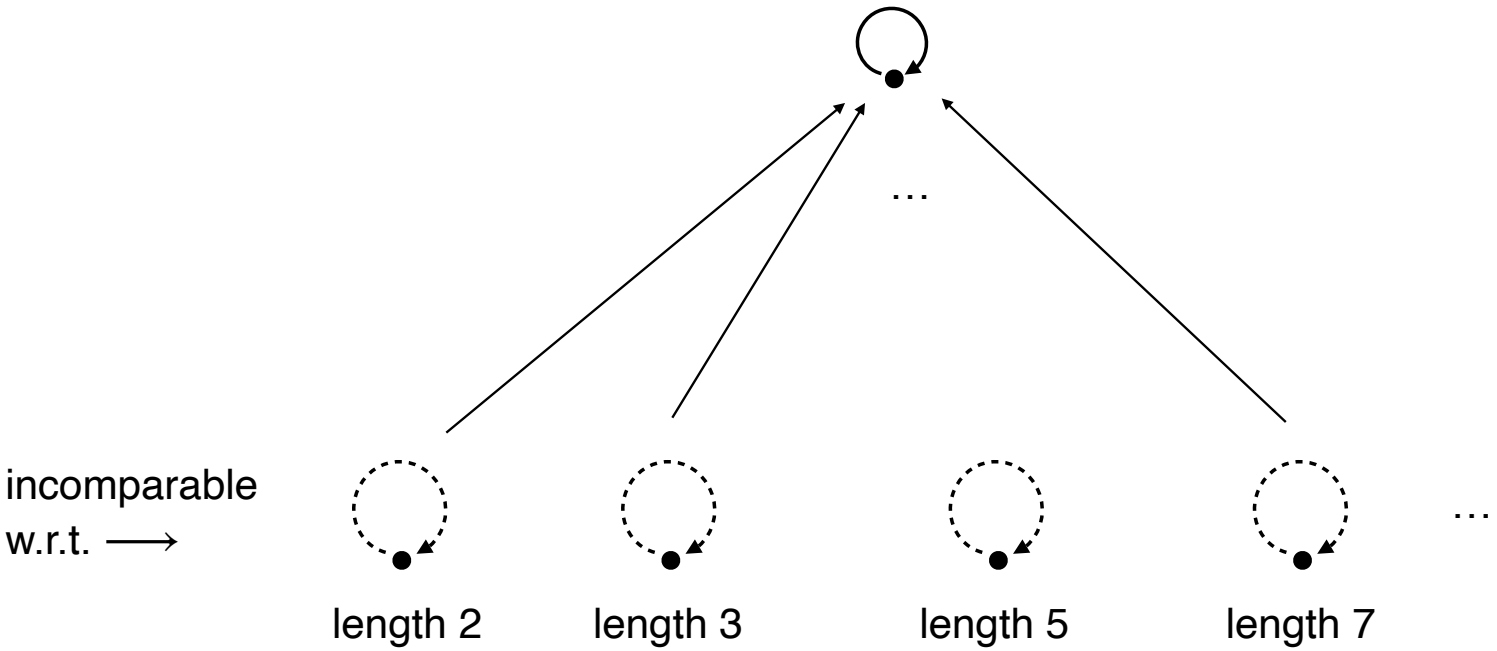
     incidence graph acyclic (aka Bergé-acyclicity)

     it actually suffices that homomorphism core of q_2 has acyclic connected component

  2. these are the only density gaps

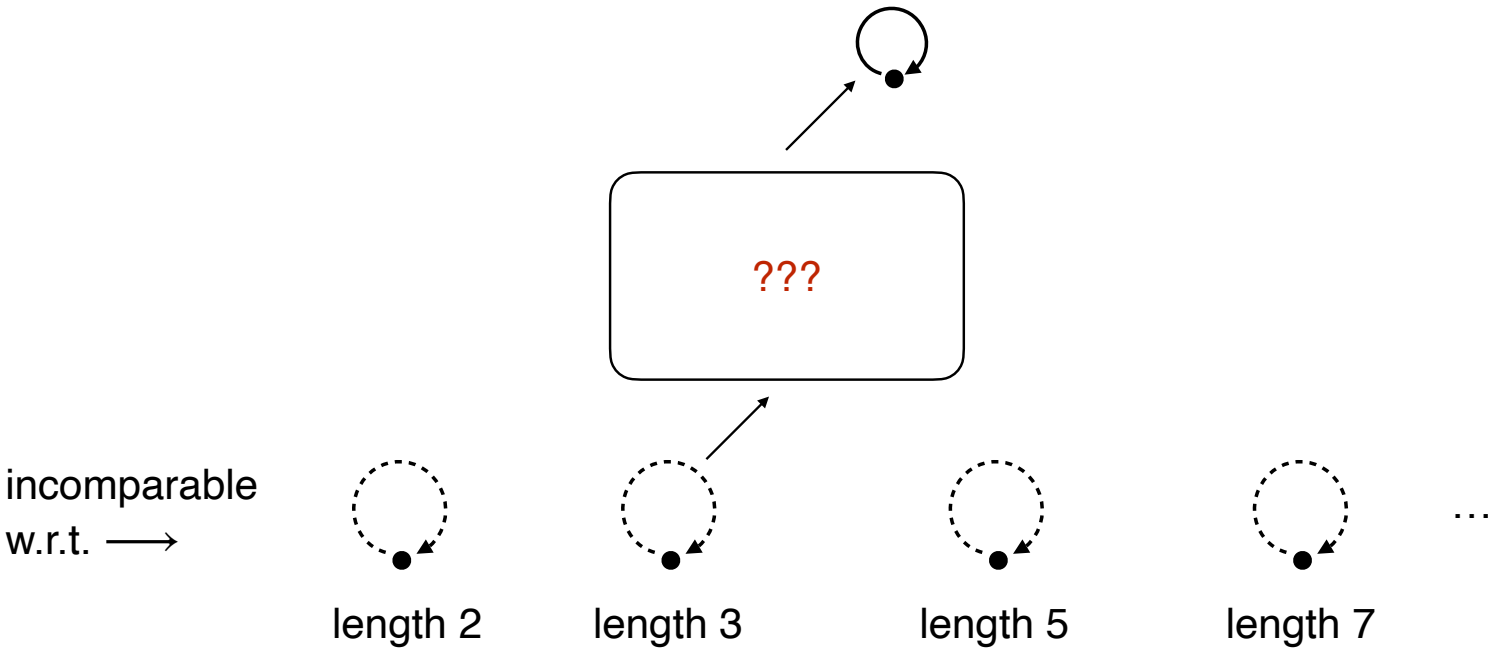
The Homomorphism Lattice

A tiny glimpse (single binary relation):



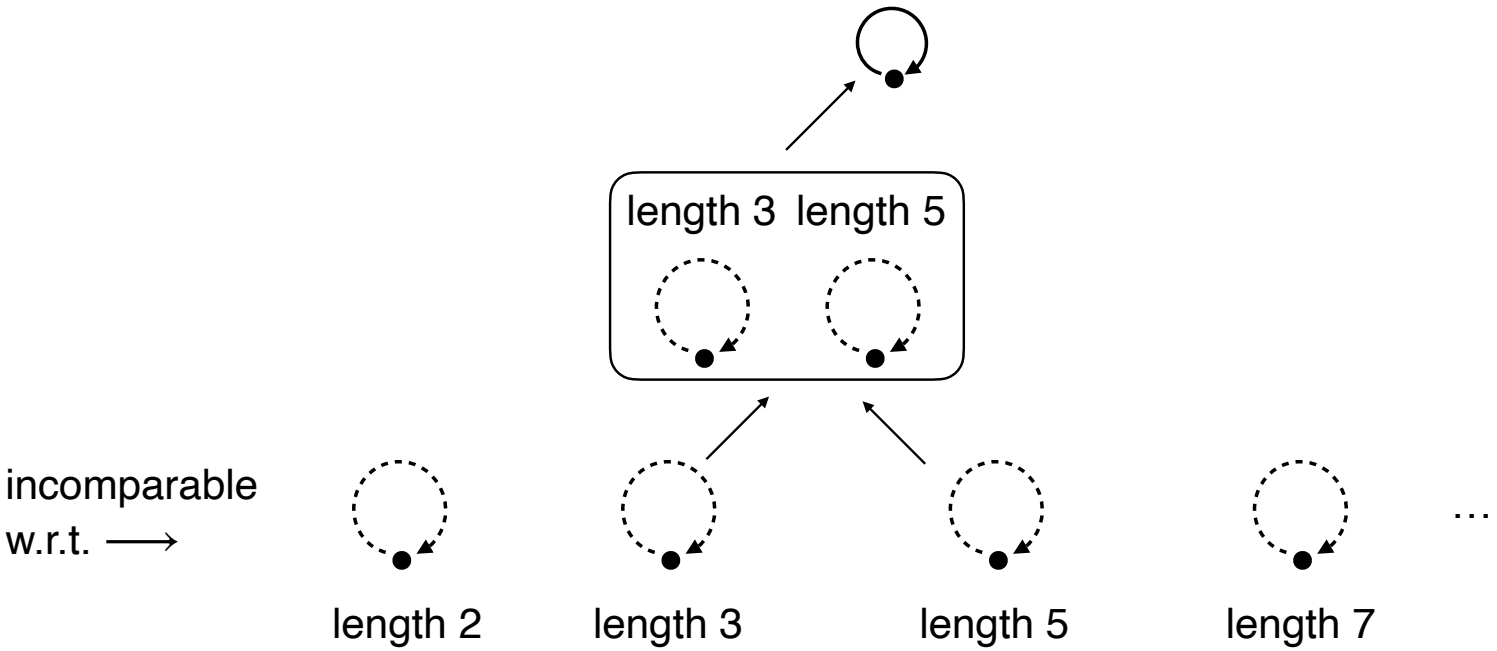
The Homomorphism Lattice

A tiny glimpse (single binary relation):



The Homomorphism Lattice

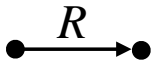
A tiny glimpse (single binary relation):



Extremal Fitting CQs

In general, there may be many different CQs that fit given set of examples, e.g.

Positive example 

Negative example 

What fitting (Boolean) CQs can you think of?

Do extremal fittings exist? Are they unique?

Are there only finitely many?

How to compute them?

How to verify them?

We can compare fittings q_1, q_2 by query containment:

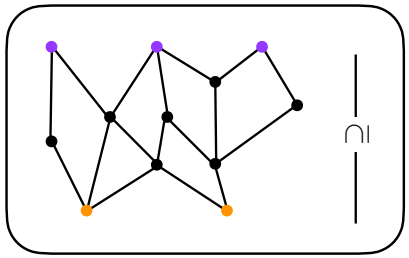
$q_1 \subseteq q_2$ means "q_2 is more general than q_1"

and "q_1 is more specific than q_2"

Two extremes:

- most-general fitting queries
  (no strictly more general query fits)
- most-specific fitting queries
  (no strictly more specific query fits)

Together describe space of all fittings:



In machine learning, this is called version space representation

Most Specific Fitting CQs

We already know everything to construct most-specific fitting CQs:

take any CQ $q(\bar{x})$ that fits set of examples E

then $q(\bar{x}) \longrightarrow (D, \bar{a})$ for every positive example $(D, \bar{a}, +) \in E$

thus $q(\bar{x}) \longrightarrow (P, \bar{b}) = \prod_{(D,\bar{a},+)\in E} (D, \bar{a})$ thus $q_{(P,\bar{b})} \subseteq q$

$(P, \bar{b})$ viewed as CQ

It follows that:

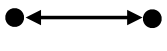
- a most-specific fitting CQ always exists (if there is any fitting CQ at all)
- it is unique (up to equivalence), more precisely: it is $q_{(P,\bar{b})}$
- Fitting Existence is coNExpTime-complete, Fitting Construction is in ExpTime

Intuition: we stick to the positive examples as closely as possible

Side Remark: For acyclic CQs, the situation is different (e.g. existence not guaranteed)

Most General Fitting CQs

Most-general fitting CQs need not exist, even when there is a fitting CQ

Consider database $K_2 =$ 

Graph G has homomorphism to K_2

iff G is 2-colorable

iff G has no cycle of odd length

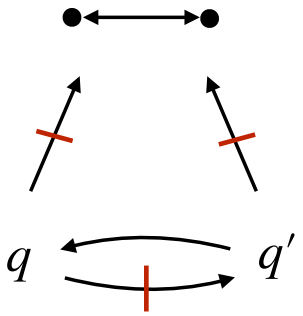
Consider following set of examples:

- no positive example
- negative example K_2

Let q be any fitting. We show: q is not most general

q must contain odd cycle, say of length k

let q' be the odd cycle of length $3k$



When do most general fitting CQs exist?

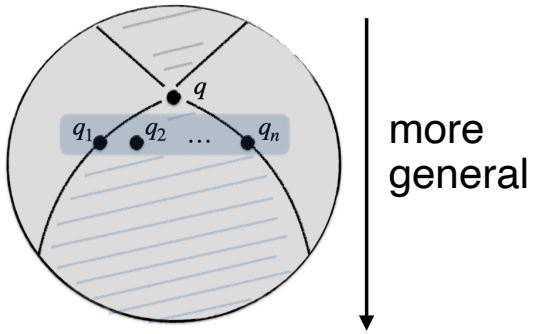
Frontiers

A frontier of a CQ is a finite complete set of minimal generalizations

Definition (Frontier)

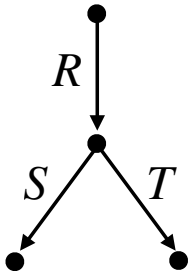
Let q be a CQ. A frontier for q is a finite set $\{q_1, \ldots, q_n\}$ such that

1. $q \subseteq q_i$ and $q_i \not\subseteq q$ for $1 \leq i \leq n$.
2. for all CQs q' with $q \subseteq q'$ and $q' \not\subseteq q$: $q_i \subseteq q'$ for some i.

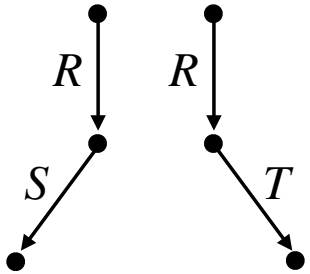


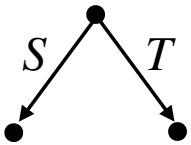
Example (Boolean):

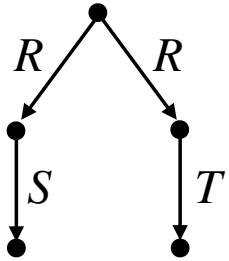
q:



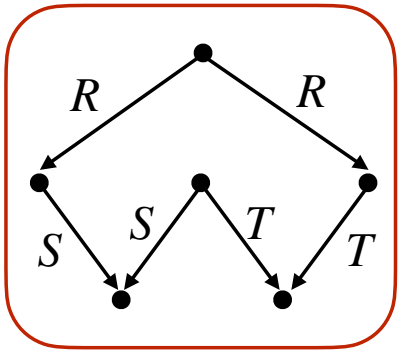
Some generalizations:







Frontier!



When do Frontiers exist?

A Boolean CQ q having a frontier implies a density gap below q in the hom-lattice

For non-Boolean CQs, we need slight generalization of acyclicity:

a CQ $q(\bar{x})$ is c-acyclic if every cycle in incidence graph passes through variable from $\bar{x}$

Theorem [tenCateDalmau2021]

1. A CQ has a frontier if and only if its homomorphism core is c-acyclic.
2. The frontier of a c-acyclic CQ can be computed in polynomial time.

Point 1 is essentially a consequence of the mentioned results on density of hom-lattice

We sketch the construction underlying Point 2

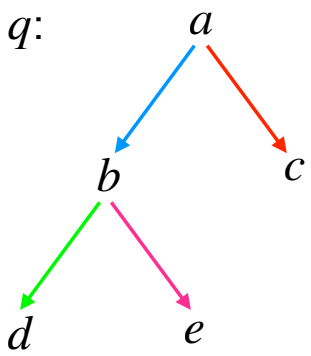
For simplicity, we consider only Boolean acyclic connected CQs

Frontiers

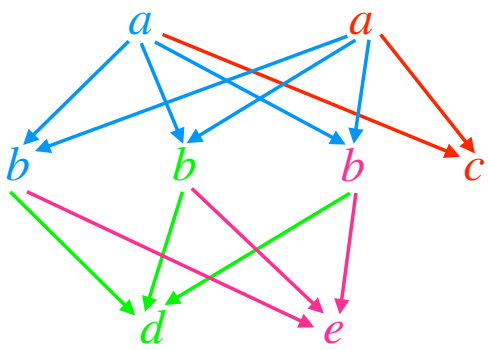
Given a Boolean acyclic connected CQ q (wlog assume to be core), do the following:

- introduce copies of each variable, one for each atom in which it occurs
- link the copies exactly like the original variables
- drop any edge between two copies that are both associated with that edge

q:



(Only) CQ in frontier:

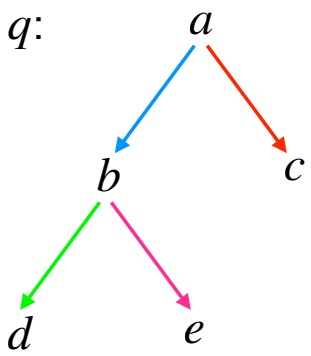


Frontiers

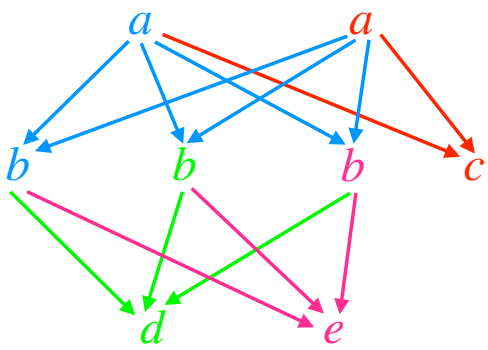
Given a Boolean acyclic connected CQ q (wlog assume to be core), do the following:

- introduce copies of each variable, one for each atom in which it occurs
- link the copies exactly like the original variables
- drop any edge between two copies that are both associated with that edge

q:



(Only) CQ in frontier:



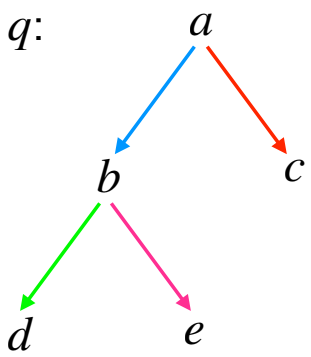
Construction clearly works in polynomial time

Frontier contains only single CQ; but no longer when Booleanness / connectedness is dropped!

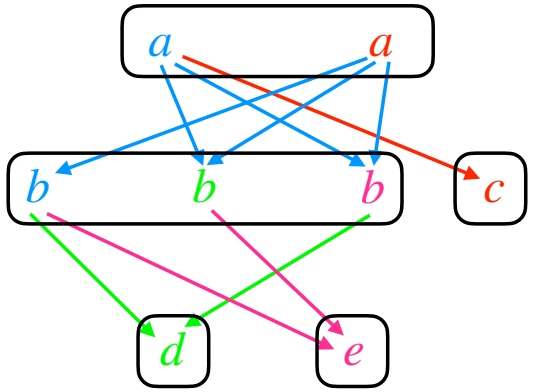
Frontiers

Given a Boolean acyclic connected CQ q (wlog assume to be core), do the following:

- introduce copies of each variable, one for each atom in which it occurs
- link the copies exactly like the original variables
- drop any edge between two copies that are both associated with that edge



(Only) CQ in frontier:



Construction clearly works in polynomial time

Frontier contains only single CQ; but no longer when Booleanness / connectedness is dropped!

Note: CQs in frontier are not trees, but still close to trees

Most-General Fitting CQs

Characterization of most-general fitting CQs:

Proposition

Let E be a collection of examples and q a CQ. TFAE:

1. q is a most-general fitting for E
2. (i) q fits E, (ii) q has a frontier F, (iii) every element of F has a homomorphism to a negative example in E

Intuition: we stick to the negative examples as closely as possible

(if we generalize q just a tiny little bit, we lose a negative example)

Relevant consequence:

If a collection E of examples admits a most-general fitting CQ q, then

q is equivalent to a c-acyclic CQ!

Most-General Fitting CQs - Verification

Theorem

Verifying whether a given CQ is a most-general fitting for a given set of examples E is NP-complete.

Easier than unrestricted fitting verification, which is DP-complete

"In NP": Given a CQ q and set of examples E, do the following

- Verify that q is equivalent to a c-acyclic CQ:

  - q is equivalent to c-acyclic CQ iff it is equivalent to c-acyclic CQ not larger than q
  - we may thus guess c-acyclic q' and homomorphisms showing $q \longrightarrow q' \longrightarrow q$

- Verify that q' fits E — possible in polytime since q' is c-acyclic
- Compute frontier F of q' (in polytime) and verify that every $\widehat{q} \in F$ has homomorphism to negative example (guess it)

Most-General Fitting CQs - Verification

Theorem

Verifying whether a given CQ is a most-general fitting for a given set of examples E is NP-complete.

Easier than unrestricted fitting verification, which is DP-complete

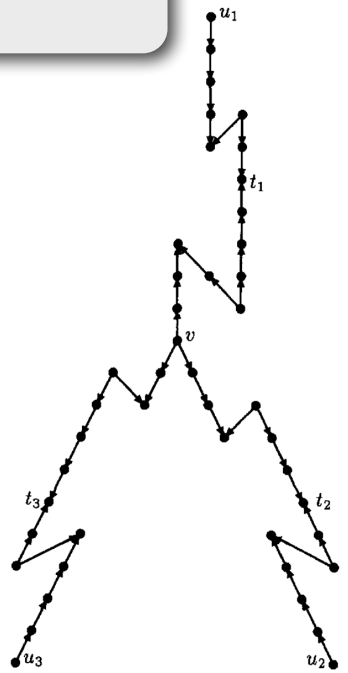
"NP-hard":

For a fixed directed graph G, let CSP(G) be the problem to decide, given a directed graph G', whether $G' \longrightarrow G$.

It is known that there is a tree T such that CSP(T) is NP-complete.

Since T is a tree, it has a frontier F. Consider $E = \{(D,-) \mid D \in F\}$

Then $G \in$ CSP(T) iff $T \uplus G$ (viewed as CQ) is a most-general fitting for E



Most-General Fitting CQs - Existence

Theorem

Given a set of examples E, deciding whether E admits a most-general fitting is ExpTime-complete

Easier than unrestricted fitting existence, which is coNExpTime-complete

Most-general fittings are c-acyclic and

c-acyclic CQs can be encoded as node-labeled trees over a finite alphabet

We can build tree automaton $\mathscr{A}$ that

- takes (encoded) c-acyclic CQ q as input
- verifies that q fits E
- verifies that every element of the frontier of q has a homomorphism to some negative example

(recall: "close to trees")

Thus $L(\mathscr{A}) = \varnothing$ iff E admits no most-general fitting

Automaton has single exponentially many states, emptiness can be checked in polytime

Most-General Fitting CQs

Most-general fitting CQs need not be unique

Consider following set of examples:

- no positive examples
- three negative examples: $\{P(a)\}$ $\{Q(a)\}$ $\{R(a)\}$

Most-general fitting CQs (Boolean):

$$\exists x \exists y\, P(x) \wedge Q(y) \qquad \exists x \exists y\, Q(x) \wedge R(y) \qquad \exists x \exists y\, P(x) \wedge R(y)$$

Definition (Basis of Most-General Fitting CQs)

A finite set B of CQs is a basis of most-general fitting CQs for E if

1. every CQ in B fits E
2. for every CQ q that fits E, there is a $\widehat{q} \in B$ with $q \subseteq \widehat{q}$ (that is: $\widehat{q} \longrightarrow q$)

How can we verify / decide the existence of minimal bases of most-general fitting CQs?

Homomorphism Duality

Definition (Homomorphism Duality)

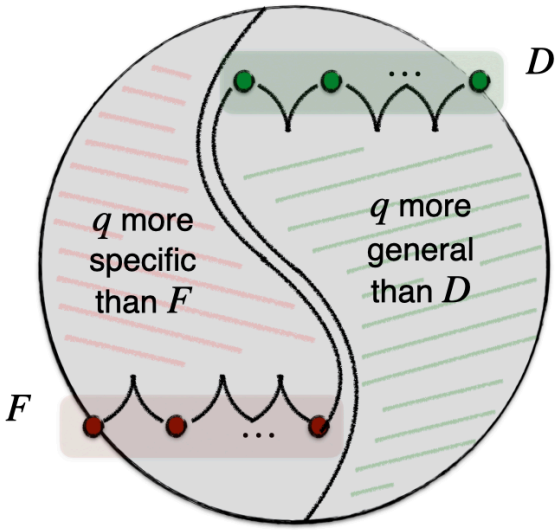
A pair (F, D), with F, D sets of CQs, is a homomorphism duality if for all CQs q, TFAE:

1. $\widehat{q} \subseteq q$ for some $\widehat{q} \in D$ (that is: $q \longrightarrow \widehat{q}$)
2. $q \not\subseteq \widehat{q}$ for all $\widehat{q} \in F$ (that is: $\widehat{q} \not\longrightarrow q$)

Partitions space of all CQs into two sets: CQs that

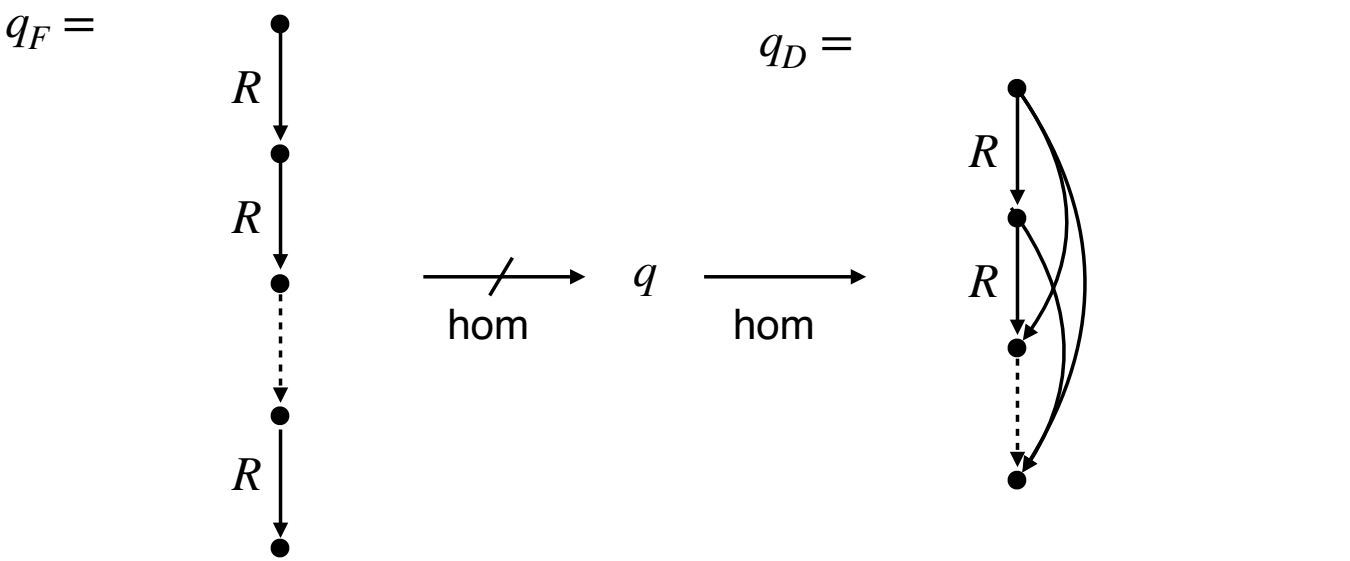
- admit a homomorphism to some CQ in D (green part)
  $\approx$ are more general than some CQ in D
- admit a homomorphism from some CQ in F (red part)
  $\approx$ are more specific than some CQ in F

F for "forbidden patterns", D for "dual"



Homomorphism Duality

Let $F = \{q_F\}$, with



path of length k

transitive tournament of length $k-1$

(F, D) is homomorphism duality for $D = \{q_D\}$

Intuitively: q_D is maximally "homomorphically strong" while avoiding q_F

Homomorphism Duality

Theorem [AlexeEtAl2011]

A CQ q participates in a homomorphism duality $(\{q\}, D)$ if and only if the homomorphism core of q is c-acyclic.

Similarly for dualities (F, D), but we have to be careful about redundancies

In contrast to the case of frontiers, D may be exponentially large (with single query)

[NesetrilTardif2005]

There is close connection between frontiers and homomorphism dualities:

- if $(\{q\}, D)$ is homomorphism duality, then $\{q \times \hat{q} \mid \hat{q} \in D\}$ is frontier for q
- conversely, from frontier for q we can construct homomorphism duality $(\{q\}, D)$

Bases and Dualities

We want characterization of finite bases of most-general fittings in terms of dualities

Proposition

Let E be a collection of negative examples and Q a finite set of CQs. TFAE:

1. Q is a basis of most-general fittings for E
2. (Q, Q_E) is homomorphism duality, $Q_E = \{q_{(D,\bar{a})} \mid (D, \bar{a}, -) \in E\}$

This is quite intuitive:

- every fitting CQ must admit a homomorphism from some CQ in Q
- no fitting CQ must admit a homomorphism to any negative example

Bases and Dualities

We want characterization of finite bases of most-general fittings in terms of dualities

Proposition

Let E be a collection of negative examples and Q a finite set of CQs. TFAE:

1. Q is a basis of most-general fittings for E
2. (Q, Q_E) is homomorphism duality, $Q_E = \{q_{(D,\bar{a})} \mid (D, \bar{a}, -) \in E\}$

Verification of bases of most-general fittings: (only negative examples)

HomDual Verification, that is, given (F, D) decide whether it is a homomorphism duality

NP-hard and in ExpTime

Existence of bases of most-general fittings:

HomDual Existence, that is, given D decide whether there is F such that (F, D) is hom. duality

NP-complete (upper bound quite non-trivial [LaroseLotenTardif07]) (only negative examples)

Relativized Dualities

To include positive examples, we need relativized form of homomorphism duality

Definition (Relativized Homomorphism Duality)

A pair (F, D), with F, D sets of CQs is a homomorphism duality relative to a CQ q_0 if for all CQs q with $q_0 \subseteq q$, TFAE:

1. $\widehat{q} \subseteq q$ for some $\widehat{q} \in D$
2. $q \not\subseteq \widehat{q}$ for all $\widehat{q} \in F$

Proposition

Let E be a collection of ~~negative~~ examples and Q a finite set of CQs. TFAE:

1. Q is a basis of most-general fittings for E
2. each $q \in Q$ fits the positive examples in E and (Q, Q_E^-) is a homomorphism duality relative to $q_0 =$ where $Q_E^- = \{q_{(D,\bar{a})} \mid (D, \bar{a}, -) \in E\}$

Relativized Dualities

Generalizing a construction of Briceno, Bulatov, Dalmau, Larose [2021] yields:

Theorem

Relativized HomDual Existence is NP-complete

Theorem

Deciding whether a given set of examples E admits a finite basis of most-general fitting CQs is NExpTime-complete.

For the verification of bases of most-general fittings, a careful analysis shows:

Theorem

Verifying whether a given finite set of CQs is a basis of most-general fittings for a given set of examples E is NExpTime-complete.

Summary of Complexity Results

	Verification	Existence	Construction
Any Fitting	DP-c	coNExpTime-c	In ExpTime
Most-Specific	NExpTime-c	coNExpTime-c	In ExpTime
Most-General	NP-c	ExpTime-c	In 2ExpTime
Basis of Most-General	NExpTime-c	NExpTime-c	In 3ExpTime
Unique	NExpTime-c	NExpTime-c	In ExpTime

Also have results for unions of CQs (UCQs) and for tree-shaped CQs

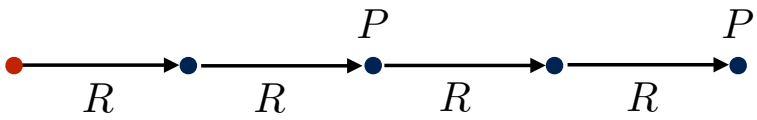
Are CQs PAC Learnable? Efficiently so?

Some basic observations:

Recent survey [tenCateFunkJungL__24]

- general and (very) classic results from learning theory (Blumer et al. 89) imply:
  CQs can be PAC learned with linear sample complexity, i.e.:
  number of required training examples depends only linearly on
  desired success probability, desired maximum error, and target CQ size
- but this is not possible in (randomized) polynomial time unless RP = NP [Kietz93]

Possibly more surprising:
lower bound even holds for unary path queries



Theorem [Kietz93]

Unary path queries are not PAC learnable in randomized polynomial time, unless RP = NP.

Proof via NP-hardness of fitting problem

Bounded Fitting

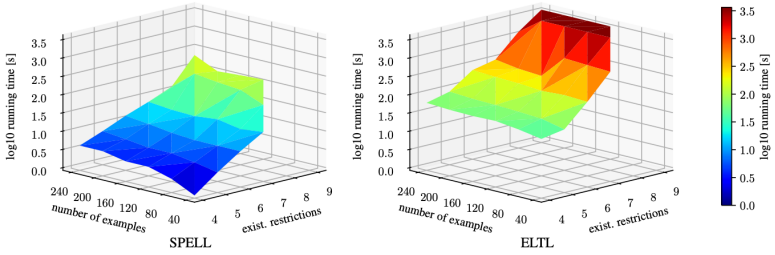
Bounded fitting approach (spirit of bounded model checking):

- try fitting CQs of increasing size $s = 1,2,3,\ldots$
- return shortest fitting CQ found (Occam algorithm, therefore PAC)

Size-bounded fitting problem:

Input: collection E of labeled examples and $s \geq 1$

Question: is there CQ q of size $\leq s$ that fits E?



Theorem

The size-bounded fitting problem is

1. Σ_2^p-complete for CQs [GottlobLeoneScarcello97]
2. NP-complete for CQs of treewidth bounded by some constant k

In Case 2, one can use a SAT solver SPELL system (2023)

Power to the Learner!

We can get to polynomial time by giving learner access to "membership oracle":

learner can present example database to oracle and ask for label (positive/negative)

Theorem

CQs are PAC learnable in polynomial time (and with linear sample complexity) using membership oracles.

Main idea:

- let positive examples be $(D_1, \bar{a}_1, +), \ldots, (D_n, \bar{a}_n, +)$ (negative examples ignored)
- start with hypothesis CQ $q_{D_1, \bar{a}_1}$, then for $i = 2, \ldots, n$:
  1. take product with $(D_i, \bar{a}_i)$ and 2. minimize hypothesis CQ using membership queries

This yields a polynomial time algorithm because of the minimization and

a PAC algorithm because we return a CQ no larger than the target CQ (Occam!)