Updated 2020/2/9

T2: Complexity of Query Evaluation
L9: Query containment & Homomorphisms

Wolfgang Gatterbauer

CS7240 Principles of scalable data management (sp20)

https://northeastern-datalab.github.io/cs7240/sp20/

Date: 2020/2/4

Outline: Complexity of Query Equivalence

- Query equivalence and query containment
  - Graph homomorphisms
  - Homomorphism beyond graphs
  - CQ containment
  - Beyond CQs
  - CQ equivalence under bag semantics
  - CQ minimization
  - Nested queries
  - Tree pattern queries

Query Equivalence

Two queries q_1, q_2 are equivalent, denoted $q_1 \equiv q_2$, if for every database instance D, we have $q_1(D) = q_2(D)$.

Query q_1 is contained in query q_2 , denoted $q_1 \subseteq q_2$, if for every database instance D, we have $q_1(D) \subseteq q_2(D)$

Boolean

$q_1 \Rightarrow q_2$

<u>Corollary</u>

$q_1 \equiv q_2$ is equivalent to ($q_1 \subseteq q_2$ and $q_1 \supseteq q_2$)

If queries are Boolean, then query containment = logical implication:

$q_1 \Leftrightarrow q_2$ is equivalent to ($q_1 \Rightarrow q_2$ and $q_1 \Leftarrow q_2$)

Homomorphisms

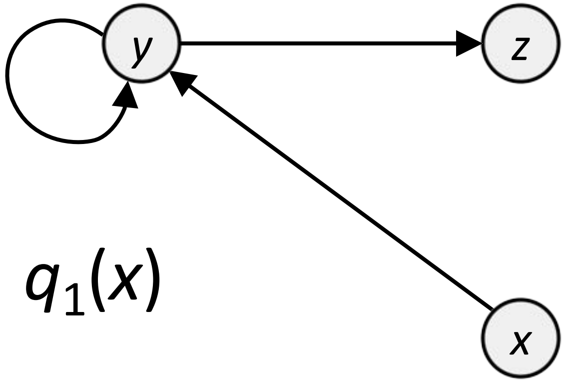
A homomorphism h from Boolean q_2 to q_1 is a function
h: var(q_2) → var(q_1) ∪ const(q_1) such that:
for every atom $R(x_1,x_2,\ldots)$ in q_2, there is an atom $R(h(x_1), h(x_2), \ldots)$ in q_1

SAME

<u>Example</u>

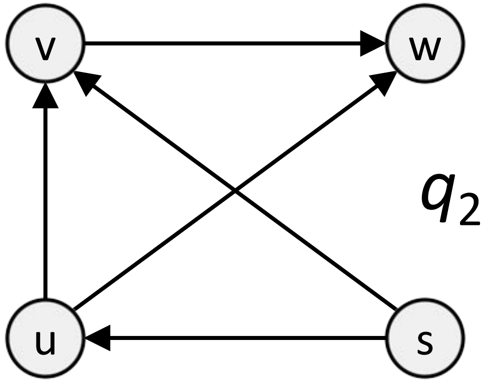
$q_1(x)$:- $R(x,y), R(y,y), R(y,z)$

$q_2(s)$:- $R(s,u), R(u,w), R(s,v), R(v,w), R(u,v)$



$q_1(x)$

$h_{2\to 1}$: ?



$q_2(x)$

Homomorphisms

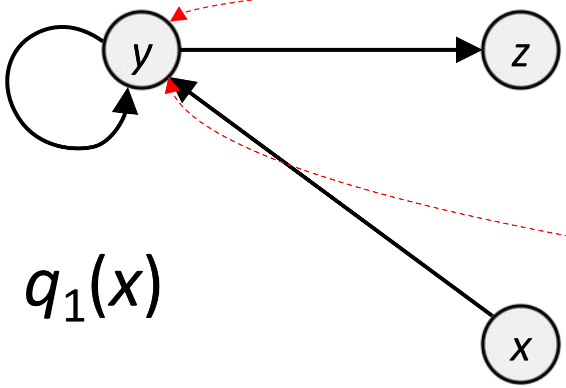
A homomorphism h from Boolean q_2 to q_1 is a function
h: var(q_2) → var(q_1) ∪ const(q_1) such that:

for every atom $R(x_1,x_2,\ldots)$ in q_2, there is an atom $R(h(x_1), h(x_2), \ldots)$ in q_1

<u>Example</u>

$q_1(x)$:- $R(x,y), R(y,y), R(y,z)$

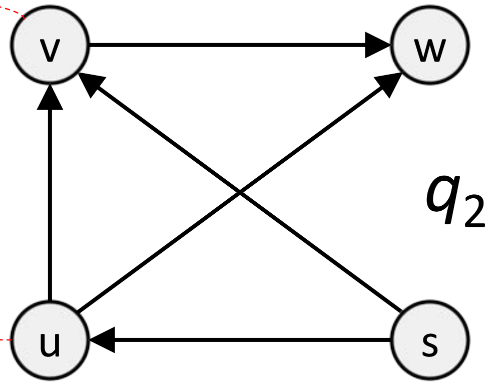
$q_2(s)$:- $R(s,u), R(u,w), R(s,v), R(v,w), R(u,v)$



$q_1(x)$

(also: $h_{2\to1}'$: {s,u,v,w} → {y})

$h_{2\to1}$: {(s,x),(u,y),(v,y),(w,z)}



$q_2(x)$

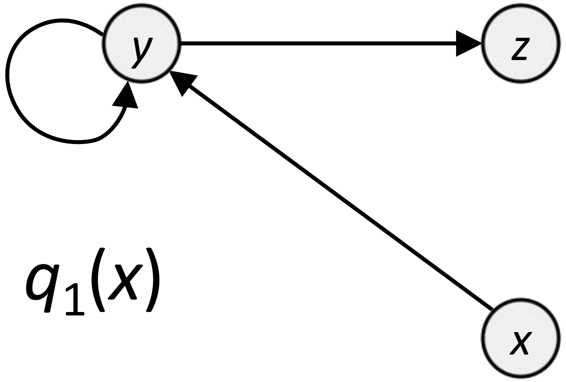
Homomorphisms

A homomorphism h from Boolean q_2 to q_1 is a function
h: var(q_2) → var(q_1) ∪ const(q_1) such that:
  for every atom $R(x_1,x_2,\ldots)$ in q_2, there is an atom $R(h(x_1), h(x_2), \ldots)$ in q_1

Example

$q_1(x)$:- $R(x,y)$, $R(y,y)$, $R(y,z)$

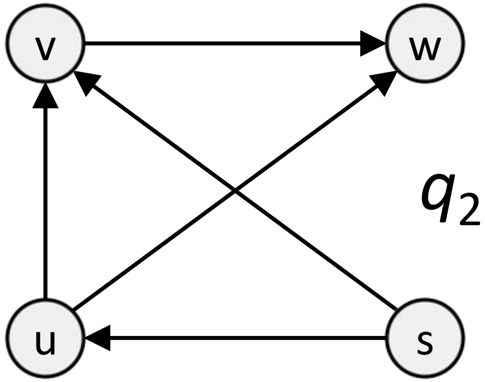
$q_2(s)$:- $R(s,u)$, $R(u,w)$, $R(s,v)$, $R(v,w)$, $R(u,v)$



$q_1(x)$

$h_{1\rightarrow 2}$: ?

$h_{2\rightarrow 1}$: $\{(s,x),(u,y),(v,y),(w,z)\}$



$q_2(x)$

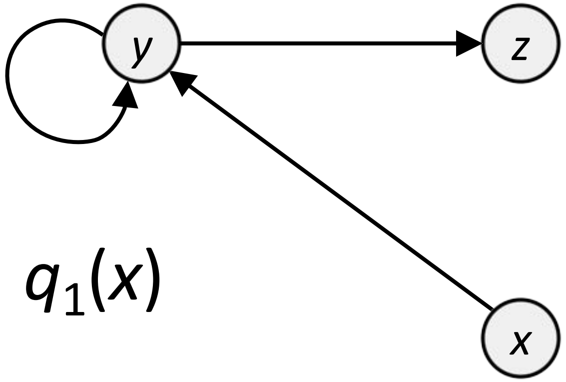
Homomorphisms

A homomorphism h from Boolean q_2 to q_1 is a function
h: var(q_2) → var(q_1) ∪ const(q_1) such that:
for every atom $R(x_1,x_2,\ldots)$ in q_2, there is an atom $R(h(x_1), h(x_2), \ldots)$ in q_1

Example

$q_1(x)$:- $R(x,y)$, $R(y,y)$, $R(y,z)$

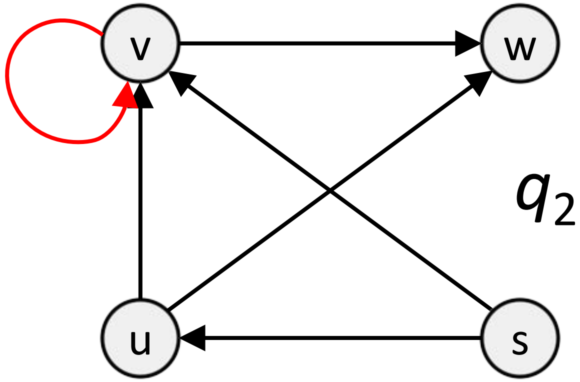
$q_2(s)$:- $R(s,u)$, $R(u,w)$, $R(s,v)$, $R(v,w)$, $R(u,v)$, $R(v,v)$



$q_1(x)$

~~$h_{1\to2}$: {(x,s),(y,v),(z,w)}~~

$h_{2\to1}$: {(s,x),(u,y),(v,y),(w,z)}



$q_2(x)$

Homomorphisms

A homomorphism h from Boolean q_2 to q_1 is a function
h: var(q_2) → var(q_1) ∪ const(q_1) such that:
for every atom $R(x_1,x_2,...)$ in q_2, there is an atom $R(h(x_1), h(x_2), ...)$ in q_1

P(2,2)

Compare to our earlier example:

$\exists x.\ P(x,x) \Rightarrow \exists x.\exists y.\ P(x,y)$

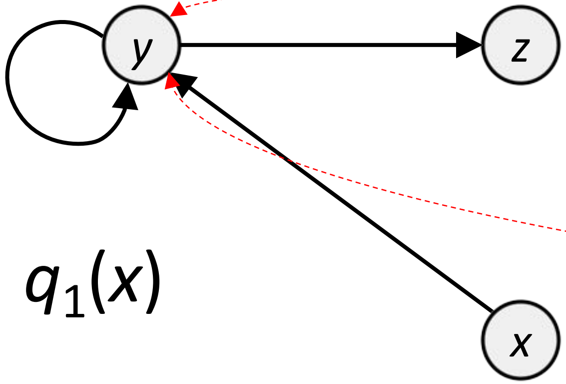
F T

P(1,2)

<u>Example</u>

$q_1(x)$:- $R(x,y), R(y,y), R(y,z)$

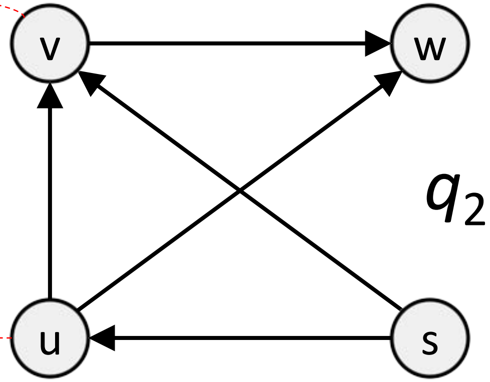
$q_2(s)$:- $R(s,u), R(u,w), R(s,v), R(v,w), R(u,v)$



$q_1(x)$

~~$h_{1\to 2}$: {(x,s),(y,v),(z,w)}~~ $q_1 \nsubseteq q_2$

$h_{2\to 1}$: {(s,x),(u,y),(v,y),(w,z)} $q_1 \subseteq q_2$



$q_2(x)$

Canonical database

Definition (Canonical database)

Given a conjunctive query q, the canonical database $D_c[q]$ is the database instance where each atom in q becomes a fact in the instance.

Example

$q_1(x)$:- $R(x,y)$, $R(y,y)$, $R(y,z)$

$D_c[q]$ = ?

Canonical database

Definition (Canonical database)

Given a conjunctive query q, the canonical database $D_c[q]$ is the database instance where each atom in q becomes a fact in the instance.

Example

$q_1(x)$:- $R(x,y)$, $R(y,y)$, $R(y,z)$

$D_c[q]$ = {R('x','y'), R('y','y'), R('y','z')}

≡ {R(a,b), R(b,b), R(b,c)}

Just treat each variable as different constant ☺

[Chandra and Merlin 1977]

THEOREM (Query Containment)

Given two Boolean CQs q_1, q_2, the following statements are equivalent:

1) $q_1 \subseteq q_2$ $\quad q_1 \Rightarrow q_2$
2) There is a homomorphism $h_{2\rightarrow 1}$ from q_2 to q_1
3) $q_2(D_C[q_1])$ is true

We will only look at 2) ⇒ 1)

[Chandra and Merlin 1977]

If there is a homomorphism h *from* q_2 *to* q_1 , then $q_1 \subseteq q_2$

1. Given $h = h_{2\rightarrow1}$, we will show that for any D: q_1(D) $\Rightarrow$ q_2(D)
2. For q_1(D) to hold, there is a valuation v s.t. $v(q_1) \in$ D
3. We will show that the composition $g = v \circ h$ is a valuation for q_2

$g = v \circ h$

$g(x) = v(h(x))$

[Chandra and Merlin 1977]

If there is a homomorphism h *from* q_2 *to* q_1, then $q_1 \subseteq q_2$

1. Given $h=h_{2\rightarrow 1}$, we will show that for any D: q_1(D) $\Rightarrow$ q_2(D)
2. For q_1(D) to hold, there is a valuation v s.t. $v(q_1) \in$ D
3. We will show that the composition $g = v \circ h$ is a valuation for q_2
   3a. By definition of h, for every $R(x_1,x_2,\ldots)$ in q_2, $R(h(x_1),h(x_2),\ldots)$ in q_1
   3b. By definition of v, for every $R(x_1,x_2,\ldots)$ in q_2, $R(v(h(x_1)),v(h(x_2)),\ldots)$ in D

$g=v \circ h$
$g(x)=v(h(x))$

[Chandra and Merlin 1977]

If there is a homomorphism h *from* q_2 *to* q_1, then $q_1 \subseteq q_2$

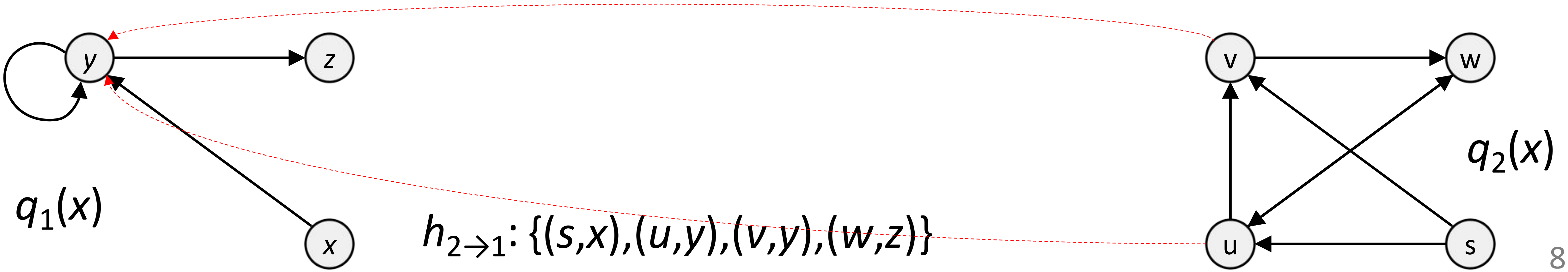
1. Given $h = h_{2\to1}$, we will show that for any D: $q_1(D) \Rightarrow q_2(D)$
2. For $q_1(D)$ to hold, there is a valuation v s.t. $v(q_1) \in D$
3. We will show that the composition $g = v \circ h$ is a valuation for q_2
   - 3a. By definition of h, for every $R(x_1, x_2, \ldots)$ in q_2, $R(h(x_1), h(x_2), \ldots)$ in q_1
   - 3b. By definition of v, for every $R(x_1, x_2, \ldots)$ in q_2, $R(v(h(x_1)), v(h(x_2)), \ldots)$ in D

$g = v \circ h$
$g(x) = v(h(x))$

<u>Example</u>

$q_1()$:- $R(x,y), R(y,y), R(y,z)$

$q_2()$:- $R(s,u), R(u,w), R(s,v), R(v,w), R(u,v)$



[Chandra and Merlin 1977]

If there is a homomorphism h *from* q_2 *to* q_1, then $q_1 \subseteq q_2$

1. Given $h = h_{2\to1}$, we will show that for any D: $q_1(D) \Rightarrow q_2(D)$
2. For $q_1(D)$ to hold, there is a valuation v s.t. $v(q_1) \in D$
3. We will show that the composition $g = v \circ h$ is a valuation for q_2
   - 3a. By definition of h, for every $R(x_1,x_2,\ldots)$ in q_2, $R(h(x_1),h(x_2),\ldots)$ in q_1
   - 3b. By definition of v, for every $R(x_1,x_2,\ldots)$ in q_2, $R(v(h(x_1)),v(h(x_2)),\ldots)$ in D

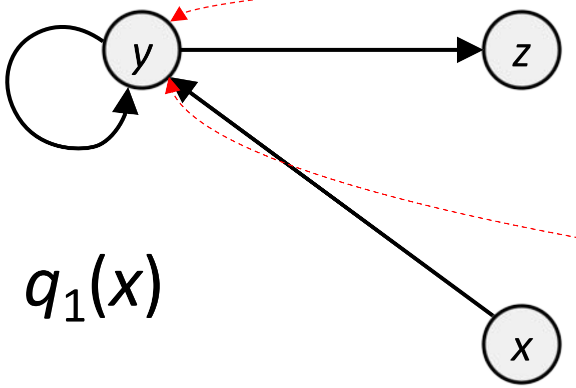
$g = v \circ h$

$g(x) = v(h(x))$

Example

$q_1()$:- $R(x,y), R(y,y), R(y,z)$

$q_2()$:- $R(s,u), R(u,w), R(s,v), R(v,w), R(u,v)$

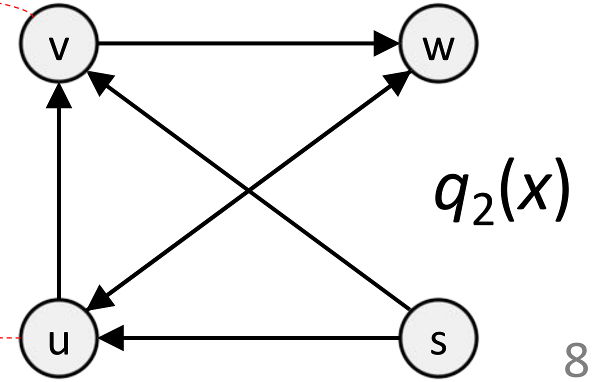


$v = \{(x,a),(y,b),(z,c)\}$

R

A	B
a	b
b	b
b	c

$h_{2\to1}$: $\{(s,x),(u,y),(v,y),(w,z)\}$



[Chandra and Merlin 1977]

If there is a homomorphism h *from* q_2 *to* q_1, then $q_1 \subseteq q_2$

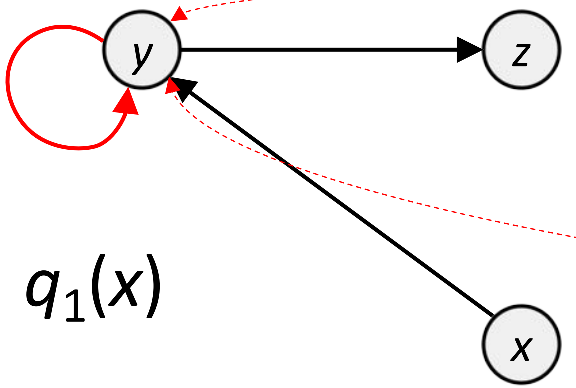
1. Given $h=h_{2\rightarrow 1}$, we will show that for any D: $q_1(\text{D}) \Rightarrow q_2(\text{D})$
2. For $q_1(\text{D})$ to hold, there is a valuation v s.t. $v(q_1) \in \text{D}$
3. We will show that the composition $g = v \circ h$ is a valuation for q_2
   3a. By definition of h, for every $R(x_1,x_2,\ldots)$ in q_2, $R(h(x_1),h(x_2),\ldots)$ in q_1
   3b. By definition of v, for every $R(x_1,x_2,\ldots)$ in q_2, $R(v(h(x_1)),v(h(x_2)),\ldots)$ in D

$g=v \circ h$
$g(x)=v(h(x))$

<u>Example</u>

$q_1()$:- $R(x,y)$, $R(y,y)$, $R(y,z)$

$q_2()$:- $R(s,u)$, $R(u,w)$, $R(s,v)$, $R(v,w)$, $R(u,v)$



$q_1(x)$

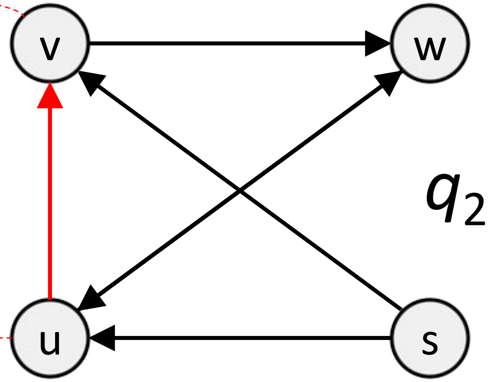
$v=\{(x,a),(y,b),(z,c)\}$

$g=\{(s,a),(u,b),(v,b),(w,c)\}$

$h_{2\rightarrow 1}$: $\{(s,x),(u,y),(v,y),(w,z)\}$

R

A	B
a	b
b	b
b	c



$q_2(x)$

Combined complexity of CQC and CQE

Corollary:

The following problems are NP-complete:

1) Given two (Boolean) conjunctive queries Q and Q’, is $Q \subseteq Q'$?

2) Given a Boolean conjunctive query Q and an instance D, does $D \models Q$?

Proof:

(a) Membership in NP follows from the Homom. Theorem:

$Q \subseteq Q'$ if and only if there is a homomorphism $h: Q' \rightarrow Q$

(b) NP-hardness follows from 3-Colorability:

G is 3-colorable if and only if $Q^{K_3} \subseteq Q^{G}$.

The Complexity of Database Query Languages

	Relational Calculus	CQs
Query Eval.: Data Complexity	In LOGSPACE (hence, in P)	In LOGSPACE (hence, in P)
Query Eval.: Combined Compl.	PSPACE-complete	NP-complete
Query Equivalence & Containment	Undecidable	NP-complete

Source: Phokion Kolaitis

Outline: Complexity of Query Equivalence

- Query equivalence and query containment
  - Graph homomorphisms
  - Homomorphism beyond graphs
  - CQ containment
  - Beyond CQs
  - CQ equivalence under bag semantics
  - CQ minimization
  - Nested queries
  - Tree pattern queries

Beyond Conjunctive Queries

- What can we say about query languages of intermediate expressive power between conjunctive queries and the full relational calculus?

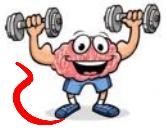
- Conjunctive queries form the sublanguage of relational algebra obtained by using only cartesian product, projection, and selection with equality conditions.

- The next step would be to consider relational algebra expressions that also involve union.

Beyond Conjunctive Queries

- Definition:
  - A union of conjunctive queries (UCQ) is a query expressible by an expression of the form $q_1 \cup q_2 \cup \ldots \cup q_m$, where each q_i is a conjunctive query.
  - A monotone query is a query expressible by a relational algebra expression which uses only union, cartesian product, projection, and selection with equality condition.
- Fact:
  - Every union of conjunctive queries is a monotone query.
  - Every monotone query is equivalent to a union of conjunctive queries, but
    - the union may have exponentially many disjuncts.
- (normal form for monotone queries).
  - Monotone queries are precisely the queries expressible by relational calculus expressions using $\wedge$, $\vee$, and $\exists$ only.

Unions of CQs and Monotone Queries



Union of Conjunctive Queries (UCQ)

Given edge relation *E*(*A*,*B*), find paths of length 1 or 2

RA ? E ∪ (unnamed RA)

RC ?

Unions of CQs and Monotone Queries

Union of Conjunctive Queries (UCQ)

Given edge relation *E*(*A*,*B*), find paths of length 1 or 2

RA $E \cup \pi_{1,4}(\sigma_{2=3}(E \times E))$ (unnamed RA)

RC ?

Unions of CQs and Monotone Queries

Union of Conjunctive Queries (UCQ)

Given edge relation *E*(*A*,*B*), find paths of length 1 or 2

RA $E \cup \pi_{1,4}(\sigma_{2=3}(E \times E))$ (unnamed RA)

RC $E(x_1, x_2) \vee \exists z[E(z, x_2) \wedge E(z, x_2)]$

Unions of CQs and Monotone Queries

Union of Conjunctive Queries (UCQ)

Given edge relation *E*(*A*,*B*), find paths of length 1 or 2

RA $E \cup \pi_{1,4}(\sigma_{2=3}(E \times E))$ (unnamed RA)

RC $E(x_1, x_2) \vee \exists z[E(z, x_2) \wedge E(z, x_2)]$

Monotone Query

Assume schema R(A,B), S(A,B), T(B,C), V(B,C)

Is following query monotone ? $(R \cup S) \bowtie (T \cup V)$

Unions of CQs and Monotone Queries

Union of Conjunctive Queries (UCQ)

Given edge relation *E*(*A*,*B*), find paths of length 1 or 2

RA $E \cup \pi_{1,4}(\sigma_{2=3}(E \times E))$ (unnamed RA)

RC $E(x_1, x_2) \vee \exists z[E(z, x_2) \wedge E(z, x_2)]$

Monotone Query

Assume schema R(A,B), S(A,B), T(B,C), V(B,C)

Is following query monotone? $(R \cup S) \bowtie (T \cup V)$

Equal to a UCQ? ?

Unions of CQs and Monotone Queries

Union of Conjunctive Queries (UCQ)

Given edge relation *E*(*A*,*B*), find paths of length 1 or 2

RA $E \cup \pi_{1,4}(\sigma_{2=3}(E \times E))$ (unnamed RA)

RC $E(x_1, x_2) \vee \exists z[E(z, x_2) \wedge E(z, x_2)]$

Monotone Query

Assume schema R(A,B), S(A,B), T(B,C), V(B,C)

Is following query monotone? $(R \cup S) \bowtie (T \cup V)$

Equal to a UCQ? $(R \bowtie T) \cup (R \bowtie V) \cup (S \bowtie T) \cup (S \bowtie V)$

The Containment Problem for Unions of CQs

THEOREM [Sagiv and Yannakakis 1981]

Let $q_1 \cup q_2 \cup \cdots \cup q_{\mathrm{m}}$ *and* $q_1' \cup q_2' \cup \cdots \cup q_n'$ *be two UCQs. Then the following are equivalent:*

1) $q_1 \cup q_2 \cup \cdots \cup q_{\mathrm{m}} \subseteq q_1' \cup q_2' \cup \cdots \cup q_n'$
2) For every $i \leq m$, there is $j \leq n$ such that $q_i \subseteq q_j'$

Proof: Use the Homomorphism Theorem

1. ⇒ 2. Since $D_C[q_i] \vDash q_i$, we have that $D_C[q_i] \vDash q_1 \cup q_2 \cup \ldots \cup q_m$ hence $D_C[q_i] \vDash q'_1 \cup q'_2 \cup \ldots \cup q'_n$, hence there is some $j \leq n$ such that $D_C[q_j]$ i $\vDash q'_j$, hence (by the Homomorphism Theorem) $q_i \subseteq q'_j$.

2. ⇒ 1. This direction is obvious.

The Complexity of Database Query Languages

	Relational Calculus	CQs	UCQs
Query Eval.: Data Complexity	In LOGSPACE (hence, in P)	In LOGSPACE (hence, in P)	In LOGSPACE (hence, in P)
Query Eval.: Combined Compl.	PSPACE-complete	NP-complete	NP-complete
Query Equivalence & Containment	Undecidable	NP-complete	NP-complete

Source: Phokion Kolaitis

Monotone Queries

- Even though monotone queries have the same expressive power as unions of conjunctive queries, the containment problem for monotone queries has higher complexity than the containment problem for unions of conjunctive queries (syntax/complexity tradeoff)

- Theorem: Sagiv and Yannakakis – 1982

  The containment problem for monotone queries is Π_2^p-complete.

- Note: The prototypical Π_2^p-complete problem is $\forall\exists$SAT, i.e., the restriction of QBF to formulas of the form

$$\forall x_1 \ldots \forall x_m \exists y_1 \ldots \exists y_n \, \phi.$$

Source: Phokion Kolaitis

The Complexity of Database Query Languages

	Relational Calculus	CQs	UCQs	Monotone queries
Query Eval.: Data Complexity	In LOGSPACE (hence, in P)	In LOGSPACE (hence, in P)	In LOGSPACE (hence, in P)	In LOGSPACE (hence, in P)
Query Eval.: Combined Compl.	PSPACE-complete	NP-complete	NP-complete	NP-complete
Query Equivalence & Containment	Undecidable	NP-complete	NP-complete	Π_2^p-complete

Source: Phokion Kolaitis

Conjunctive Queries with Inequalities

- Definition: Conjunctive queries with inequalities form the sublanguage of relational algebra obtained by using only cartesian product, projection, and selection with equality and inequality (≠, <, ≤) conditions.

- Example: Q(x,y):-- E(x,z), E(z,w),E(w,y), z ≠ w, z < y.

- Theorem: (Klug – 1988, van der Meyden – 1992)
  - The query containment problem for conjunctive queries with inequalities is Π_2^p-complete.
  - The query evaluation problem for conjunctive queries with inequalities in NP-complete.

Source: Phokion Kolaitis

The Complexity of Database Query Languages

	Relational Calculus	CQs	UCQs	Monotone queries / CQs with inequalities
Query Eval.: Data Complexity	In LOGSPACE (hence, in P)	In LOGSPACE (hence, in P)	In LOGSPACE (hence, in P)	In LOGSPACE (hence, in P)
Query Eval.: Combined Compl.	PSPACE-complete	NP-complete	NP-complete	NP-complete
Query Equivalence & Containment	Undecidable	NP-complete	NP-complete	Π_2^p-complete

Source: Phokion Kolaitis

Outline: Complexity of Query Equivalence

- Query equivalence and query containment
  - Graph homomorphisms
  - Homomorphism beyond graphs
  - CQ containment
  - Beyond CQs
  - CQ equivalence under bag semantics
  - CQ minimization
  - Nested queries
  - Tree pattern queries

Following slides are from Phokion Kolaitis's talk on "Logic and databases" at "Logical structures in Computation Boot Camp", Berkeley 2016:
https://simons.berkeley.edu/talks/logic-and-databases

Logic and Databases

Phokion G. Kolaitis

UC Santa Cruz & IBM Research – Almaden

Lecture 4 – Part 1

1

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Thematic Roadmap

- ✓ Logic and Database Query Languages
  - – Relational Algebra and Relational Calculus
  - – Conjunctive queries and their variants
  - – Datalog
- ✓ Query Evaluation, Query Containment, Query Equivalence
  - – Decidability and Complexity
- ✓ Other Aspects of Conjunctive Query Evaluation
- • Alternative Semantics of Queries
  - – Bag Databases: Semantics and Conjunctive Query Containment
  - – Probabilistic Databases: Semantics and Dichotomy Theorems for Conjunctive Query Evaluation
  - – Inconsistent Databases: Semantics and Dichotomy Theorems

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Alternative Semantics

- So far, we have examined logic and databases under classical semantics:
  - The database relations are sets.
  - Tarskian semantics are used to interpret queries definable be first-order formulas.
- Over the years, several different alternative semantics of queries have been investigated. We will discuss three such scenarios:
  - The database relations can be bags (multisets).
  - The databases may be probabilistic.
  - The databases may be inconsistent.

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Sets vs. Multisets

Relation EMPLOYEE(name, dept, salary)

- Relational Algebra Expression:

  $\pi_{salary}\ (\sigma_{dept\ =\ CS}\ (\text{EMPLOYEE}))$

- SQL query:

```
SELECT   salary
FROM     EMPLOYEE
WHERE    dpt = 'CS'
```

- SQL returns a bag (multiset) of numbers in which a number may appear several times, provided different faculty had the same salary.
- SQL does not eliminate duplicates, in general, because:
  - Duplicates are important for aggregate queries (e.g., average)
  - Duplicate elimination takes nlogn time.

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Relational Algebra Under Bag Semantics

Operation	Multiplicity
Union $R_1 \cup R_2$	$m_1 + m_2$
Intersection $R_1 \cap R_2$	$\min(m_1, m_2)$
Product $R_1 \times R_2$	$m_1 \times m_2$
Projection and Selection	Duplicates are not eliminated

- R_1

A	B
1	2
1	2
2	3

- R_2

B	C
2	4
2	5

- $(R_1 \bowtie R_2)$

A	B	C
1	2	4
1	2	4
1	2	5
1	2	5

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Conjunctive Queries Under Bag Semantics

Chaudhuri & Vardi – 1993

Optimization of ***Real*** Conjunctive Queries

- Called for a re-examination of conjunctive-query optimization under bag semantics.
- In particular, they initiated the study of the containment problem for conjunctive queries under bag semantics.
- This problem has turned out to be *much more challenging* than originally perceived.

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

PROBLEMS

Problems worthy
of attack
prove their worth
by hitting back.

in: *Grooks* by Piet Hein (1905-1996)

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Query Containment Under Set Semantics

Class of Queries	Complexity of Query Containment
Conjunctive Queries	NP-complete Chandra & Merlin – 1977
Unions of Conjunctive Queries	NP-complete Sagiv & Yannakakis - 1980
Conjunctive Queries with $\neq$, $\leq$, $\geq$	Π_2^p-complete Klug 1988, van der Meyden -1992
First-Order (SQL) queries	Undecidable Trakhtenbrot - 1949

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

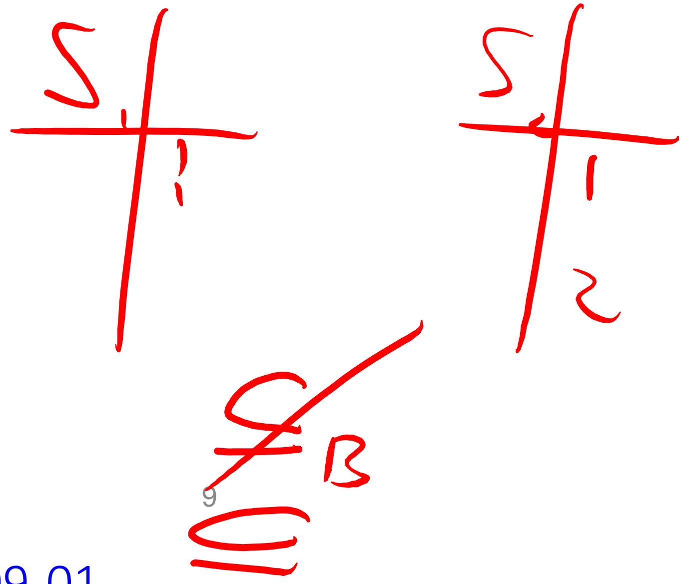
Bag Semantics vs. Set Semantics

- For bags R_1, R_2:
  $R_1 \subseteq_{BAG} R_2$ if $m(\mathbf{a},R_1) \leq m(\mathbf{a},R_2)$, for every tuple **a**.
- $Q^{BAG}(D)$: Result of evaluating Q on (bag) database D.
- $Q_1 \subseteq_{BAG} Q_2$ if for every (bag) database D, we have that

$$Q_1^{BAG}(D) \subseteq_{BAG} Q_2^{BAG}(D).$$

Fact:

- $Q_1 \subseteq_{BAG} Q_2$ implies $Q_1 \subseteq Q_2$.
- The converse does **not** always hold.



9

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Bag Semantics vs. Set Semantics

Fact: $Q_1 \subseteq Q_2$ does not imply that $Q_1 \subseteq_{BAG} Q_2$.

Example:

- $Q_1(x)$:- $P(x), T(x)$
- $Q_2(x)$:- $P(x)$

- $Q_1 \subseteq Q_2$ (obvious from the definitions)
- $Q_1 \not\subseteq_{BAG} Q_2$
- Consider the (bag) instance D = {P(a), T(a), T(a)}. Then:
  - $Q_1(D) = \{a,a\}$
  - $Q_2(D) = \{a\}$, so $Q_1(D) \not\subseteq Q_2(D)$.

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Query Containment under Bag Semantics

- Chaudhuri & Vardi - 1993 stated that:

  Under bag semantics, the containment problem for conjunctive queries is Π_2^p-hard.

- Problem:
  - What is the exact complexity of the containment problem for conjunctive queries under bag semantics?
  - Is this problem decidable?

11

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Query Containment Under Bag Semantics

- 23 years have passed since the containment problem for conjunctive queries under bag semantics was raised.
- Several attacks to solve this problem have failed.
- At least two technically flawed PhD theses on this problem have been produced.
- Chaudhuri and Vardi have withdrawn the claimed Π_2^p-hardness of this problem; no one has provided a proof.

12

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Query Containment Under Bag Semantics

- The containment problem for conjunctive queries under bag semantics remains **open** to date.

- However, progress has been made towards the containment problem under bag semantics for the two main extensions of conjunctive queries:
  - Unions of conjunctive queries
  - Conjunctive queries with $\neq$

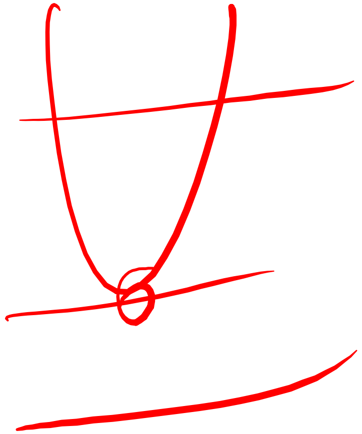
Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Unions of Conjunctive Queries

Theorem (Ioannidis & Ramakrishnan – 1995):
Under bag semantics, the containment problem for unions of conjunctive queries is **undecidable**.

Hint of Proof:
Reduction from Hilbert's 10th Problem.

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01



Hilbert's 10^{th} Problem

- Hilbert's 10^{th} Problem – 1900
  (10^{th} in Hilbert's list of 23 problems)

  Given a Diophantine equation with any number of unknown quantities and with rational integral numerical coefficients: To devise a process according to which it can be determined in a finite number of operations whether the equation is solvable in rational integers.

  In effect, Hilbert's 10^{th} Problem is:
  Find an algorithm for the following problem:
  Given a polynomial $P(x_1,\ldots,x_n)$ with integer coefficients, does it have an all-integer solution?

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Hilbert's 10^{th} Problem

- Hilbert's 10^{th} Problem – 1900
  (10^{th} in Hilbert's list of 23 problems)
  Find an algorithm for the following problem:
  Given a polynomial $P(x_1,\ldots,x_n)$ with integer coefficients, does it have an all-integer solution?

- Y. Matiyasevich – 1971
  (building on M. Davis, H. Putnam, and J. Robinson)
  - Hilbert's 10^{th} Problem is **undecidable**, hence **no** such algorithm exists.

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Hilbert's 10^{th} Problem

- Fact: The following variant of Hilbert's 10^{th} Problem is undecidable:
  - Given two polynomials $p_1(x_1,\ldots x_n)$ and $p_2(x_1,\ldots x_n)$ with positive integer coefficients and no constant terms, is it true that $p_1 \leq p_2$?

    In other words, is it true that $p_1(a_1,\ldots,a_n) \leq p_2(a_1,\ldots a_n)$, for all positive integers $a_1,\ldots,a_n$?

- Thus, there is no algorithm for deciding questions like:
  - Is $3x_1^4x_2x_3 + 2x_2x_3 \leq x_1^6 + 5x_2x_3$?

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Unions of Conjunctive Queries

Theorem (Ioannidis & Ramakrishnan – 1995):

Under bag semantics, the containment problem for unions of conjunctive queries is **undecidable**.

Hint of Proof:

- Reduction from the previous variant of Hilbert's 10^{th} Problem:
  - Use joins of unary relations to encode monomials (products of variables).
  - Use unions to encode sums of monomials.

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Unions of Conjunctive Queries

Example: Consider the polynomial $3x_1^4x_2x_3 + 2x_2x_3$

- The monomial $x_1^4x_2x_3$ is encoded by the conjunctive query $P_1(w), P_1(w), P_1(w), P_1(w), P_2(w), P_3(w)$.
- The monomial x_2x_3 is encoded by the conjunctive query $P_2(w), P_3(w)$.
- The polynomial $3x_1^4x_2x_3 + 2x_2x_3$ is encoded by the union having:
  - three copies of $P_1(w), P_1(w), P_1(w), P_1(w), P_2(w), P_3(w)$ and
  - two copies of $P_2(w), P_3(w)$.

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Complexity of Query Containment

Class of Queries	Complexity – Set Semantics	Complexity – Bag Semantics
Conjunctive queries	NP-complete CM – 1977	
Unions of conj. queries	NP-complete SY - 1980	Undecidable IR - 1995
Conj. queries with $\neq$, $\leq$, $\geq$	Π_2^p-complete vdM - 1992	
First-order (SQL) queries	Undecidable Trakhtenbrot - 1949	Undecidable

20

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Conjunctive Queries with ≠

Theorem (Jayram, K …, Vee – 2006):
Under bag semantics, the containment problem for conjunctive queries with ≠ is **undecidable**.

In fact, this problem is **undecidable** even if

- the queries use only a single relation of arity 2;
- the number of inequalities in the queries is at most some fixed (albeit huge) constant.

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Complexity of Query Containment

Class of Queries	Complexity – Set Semantics	Complexity – Bag Semantics
Conjunctive queries	NP-complete CM – 1977	**Open**
Unions of conj. queries	NP-complete SY - 1980	Undecidable IR - 1995
Conj. queries with $\neq$, $\leq$, $\geq$	Π_2^p-complete vdM - 1992	Undecidable JKV - 2006
First-order (SQL) queries	Undecidable Trakhtenbrot - 1949	Undecidable

26

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

Subsequent Developments

- Some progress has been made towards identifying special classes of conjunctive queries for which the containment problem under bag semantics is decidable.
  - Afrati, Damigos, Gergatsoulis – 2010
    - Projection-free conjunctive queries.
  - Kopparty and Rossman – 2011
    - A large class of boolean conjunctive queries on graphs.

Source: Phokion Kolaitis: https://simons.berkeley.edu/talks/phokion-kolaitis-2016-09-01

S

F

W

A (NOT EXISTS

S

F

W)

A (NOT EX

$=$ $\neq$

$\wedge$

Pointers to related work

- Kolaitis. *Logic and Databases*. Logical Structures in Computation Boot Camp, Berkeley 2016. https://simons.berkeley.edu/talks/logic-and-databases
- Abiteboul, Hull, Vianu. *Foundations of Databases*. Addison Wesley, 1995. http://webdam.inria.fr/Alice/, Ch 2.1: Theoretical background, Ch 6.2: Conjunctive queries & homomorphisms & query containment, Ch 6.3: Undecidability of equivalence for calculus.
- Chandra, Merlin. *Optimal implementation of conjunctive queries in relational data bases*. STOC 1977. https://doi.org/10.1145/800105.803397