Consistent Query Answering: The First Ten Years

Jan Chomicki
University at Buffalo

October 1, 2008

Table of contents

Integrity constraints (dependencies)

Database instance D:

- a finite first-order structure
- the information about the world

Integrity constraints (dependencies)

Database instance D:

- a finite first-order structure
- the information about the world

Integrity constraints Σ:

- first-order logic formulas
- the properties of the world

Integrity constraints (dependencies)

Database instance D:

- a finite first-order structure
- the information about the world

Integrity constraints Σ:

- first-order logic formulas
- the properties of the world

Satisfaction of constraints: $D \models \Sigma$

Formula satisfaction in a first-order structure.

Integrity constraints (dependencies)

Database instance D:

- a finite first-order structure
- the information about the world

Integrity constraints Σ:

- first-order logic formulas
- the properties of the world

Satisfaction of constraints: $D \models \Sigma$

Formula satisfaction in a first-order structure.

Consistent database: $D \models \Sigma$

Name	City	Salary
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

Integrity constraints (dependencies)

Database instance D:

- a finite first-order structure
- the information about the world

Integrity constraints Σ:

- first-order logic formulas
- the properties of the world

Satisfaction of constraints: $D \models \Sigma$

Formula satisfaction in a first-order structure.

Consistent database: $D \models \Sigma$

Name	City	Salary
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

Inconsistent database: $D \not\models \Sigma$

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

Sources of inconsistency:

- integration of independent data sources with overlapping data
- time lag of updates (eventual consistency)
- unenforced integrity constraints

▸ CQA

Whence Inconsistency?

Sources of inconsistency:

- integration of independent data sources with overlapping data
- time lag of updates (eventual consistency)
- unenforced integrity constraints

Eliminating inconsistency?

- not enough information, time, or money
- difficult, impossible or undesirable
- unnecessary: queries may be insensitive to inconsistency

▸ CQA

Whence Inconsistency?

Sources of inconsistency:

- integration of independent data sources with overlapping data
- time lag of updates (eventual consistency)
- unenforced integrity constraints

Eliminating inconsistency?

- not enough information, time, or money
- difficult, impossible or undesirable
- unnecessary: queries may be insensitive to inconsistency

Living with inconsistency?

- ignoring inconsistency
- modifying the schema
- exceptions to constraints.

▸ CQA

Ignoring Inconsistency

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

Ignoring Inconsistency

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

SELECT Name
FROM Employee
WHERE Salary $\leq$ 25M

Ignoring Inconsistency

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

SELECT Name
FROM Employee
WHERE Salary $\leq$ 25M

$\longrightarrow$

Name
Gates
Jobs

Ignoring Inconsistency

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

SELECT Name
FROM Employee
WHERE Salary $\leq$ 25M

$\longrightarrow$

Name
Gates
Jobs

Query results not reliable.

Horizontal Decomposition

Decomposition into two relations:

- violators
- the rest

(De Bra, Paredaens [DBP83])

Horizontal Decomposition

Decomposition into two relations:

- violators
- the rest

(De Bra, Paredaens [DBP83])

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name → City Salary

Decomposition into two relations:

- violators
- the rest

(De Bra, Paredaens [DBP83])

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name → City Salary

Jobs	Cupertino	10M

Name → City Salary

Gates	Redmond	20M
Gates	Redmond	30M

Name → City Salary

Weakening the contraints:

- functional dependencies $\rightsquigarrow$ denial constraints

(Borgida [Bor85])

Weakening the contraints:

- functional dependencies $\rightsquigarrow$ denial constraints

(Borgida [Bor85])

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

Weakening the contraints:

- functional dependencies $\rightsquigarrow$ denial constraints

(Borgida [Bor85])

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

$\longrightarrow$

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

except Name='Gates'

The Impact of Inconsistency on Queries

Traditional view

- query results defined irrespective of integrity constraints
- query evaluation may be optimized in the presence of integrity constraints (semantic query optimization)

The Impact of Inconsistency on Queries

Traditional view

- query results defined irrespective of integrity constraints
- query evaluation may be optimized in the presence of integrity constraints (semantic query optimization)

Our view

- inconsistency leads to uncertainty
- query results may depend on integrity constraint satisfaction
- inconsistency may be eliminated (repairing) or tolerated (consistent query answering)

Restoring consistency:

- insertion, deletion
- minimal change

Restoring consistency:

- insertion, deletion
- minimal change

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

Restoring consistency:

- insertion, deletion
- minimal change

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

Name	City	Salary
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

Name	City	Salary
Gates	Redmond	20M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

Consistent query answer:

Query answer obtained in every repair.

(Arenas, Bertossi, Ch. [ABC99])

Consistent query answer:

Query answer obtained in every repair.

(Arenas, Bertossi, Ch. [ABC99])

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

Consistent query answer:

Query answer obtained in every repair.

(Arenas, Bertossi, Ch. [ABC99])

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

SELECT Name
FROM Employee
WHERE Salary $\leq$ 25M

Name
Jobs

Consistent query answer:

Query answer obtained in every repair.

(Arenas, Bertossi, Ch. [ABC99])

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name → City Salary

SELECT Name
FROM Employee
WHERE Salary $\geq$ 10M

Name
Gates
Jobs

Formal definition

What constitutes reliable (consistent) information in an inconsistent database.

Formal definition

What constitutes reliable (consistent) information in an inconsistent database.

Algorithms

How to compute consistent information.

Research Goals

Formal definition

What constitutes reliable (consistent) information in an inconsistent database.

Algorithms

How to compute consistent information.

Computational complexity analysis

- tractable vs. intractable classes of queries and integrity constraints
- tradeoffs: complexity vs. expressiveness.

Research Goals

Formal definition

What constitutes reliable (consistent) information in an inconsistent database.

Algorithms

How to compute consistent information.

Computational complexity analysis

- tractable vs. intractable classes of queries and integrity constraints
- tradeoffs: complexity vs. expressiveness.

Implementation

- preferably using DBMS technology.

Research Goals

Formal definition

What constitutes reliable (consistent) information in an inconsistent database.

Algorithms

How to compute consistent information.

Computational complexity analysis

- tractable vs. intractable classes of queries and integrity constraints
- tradeoffs: complexity vs. expressiveness.

Implementation

- preferably using DBMS technology.

Applications

???

Repair D' of a database D w.r.t. the integrity constraints IC:

- D': over the same schema as D
- $D' \models IC$
- symmetric difference between D and D' is minimal.

Repair D' of a database D w.r.t. the integrity constraints IC:

- D': over the same schema as D
- $D' \models IC$
- symmetric difference between D and D' is minimal.

Consistent query answer to a query Q in D w.r.t. IC:

- an element of the result of Q in every repair of D w.r.t. IC.

Repair D' of a database D w.r.t. the integrity constraints IC:

- D': over the same schema as D
- $D' \models IC$
- symmetric difference between D and D' is minimal.

Consistent query answer to a query Q in D w.r.t. IC:

- an element of the result of Q in every repair of D w.r.t. IC.

Another incarnation of the idea of sure query answers [Lipski: TODS'79].

A Logical Aside

Belief revision

- semantically: repairing $\equiv$ revising the database with integrity constraints
- consistent query answers $\equiv$ counterfactual inference.

Logical inconsistency

- inconsistent database: database facts together with integrity constraints form an inconsistent set of formulas
- trivialization of reasoning does not occur because constraints are not used in relational query evaluation.

Exponentially many repairs

Example relation $R(A, B)$

- violates the dependency $A \longrightarrow B$
- has 2^n repairs.

A	B
a_1	b_1
a_1	c_1
a_2	b_2
a_2	c_2
...	
a_n	b_n
a_n	c_n

$A \longrightarrow B$

Example relation $R(A, B)$

- violates the dependency $A \longrightarrow B$
- has 2^n repairs.

A	B
a_1	b_1
a_1	c_1
a_2	b_2
a_2	c_2
...	
a_n	b_n
a_n	c_n

$A \longrightarrow B$

It is impractical to apply the definition of CQA directly.

Query Rewriting

Given a query Q and a set of integrity constraints IC, build a query Q^{IC} such that for every database instance D

> *the set of answers to Q^{IC} in D = the set of consistent answers to Q in D w.r.t. IC.*

Computing Consistent Query Answers

Query Rewriting

Given a query Q and a set of integrity constraints IC, build a query Q^{IC} such that for every database instance D

> *the set of answers to Q^{IC} in D = the set of consistent answers to Q in D w.r.t. IC.*

Representing all repairs

Given IC and D:

1. build a space-efficient representation of all repairs of D w.r.t. IC
2. use this representation to answer (many) queries.

Computing Consistent Query Answers

Query Rewriting

Given a query Q and a set of integrity constraints IC, build a query Q^{IC} such that for every database instance D

> *the set of answers to Q^{IC} in D = the set of consistent answers to Q in D w.r.t. IC.*

Representing all repairs

Given IC and D:

1. build a space-efficient representation of all repairs of D w.r.t. IC
2. use this representation to answer (many) queries.

Logic programs

Given IC, D and Q:

1. build a logic program $P_{IC,D}$ whose models are the repairs of D w.r.t. IC
2. build a logic program P_Q expressing Q
3. use a logic programming system that computes the query atoms present in all models of $P_{IC,D} \cup P_Q$.

Constraint classes

Universal constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n \vee B_1 \vee \cdots \vee B_m$

Constraint classes

Universal constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n \vee B_1 \vee \cdots \vee B_m$

Example

$\forall.\ \neg Par(x) \vee Ma(x) \vee Fa(x)$

Constraint classes

Universal constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n \vee B_1 \vee \cdots \vee B_m$

Denial constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n$

Example

$\forall.\ \neg Par(x) \vee Ma(x) \vee Fa(x)$

Constraint classes

Universal constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n \vee B_1 \vee \cdots \vee B_m$

Example

$\forall.\ \neg Par(x) \vee Ma(x) \vee Fa(x)$

Denial constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n$

Example

$\forall.\ \neg M(n, s, m) \vee \neg M(m, t, w) \vee s \leq t$

Constraint classes

Universal constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n \vee B_1 \vee \cdots \vee B_m$

Denial constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n$

Functional dependencies

$X \rightarrow Y$:

- a key dependency in F if $Y = U$
- a primary-key dependency: only one key exists

Example

$\forall.\ \neg Par(x) \vee Ma(x) \vee Fa(x)$

Example

$\forall.\ \neg M(n, s, m) \vee \neg M(m, t, w) \vee s \leq t$

Constraint classes

Universal constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n \vee B_1 \vee \cdots \vee B_m$

Denial constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n$

Functional dependencies

$X \rightarrow Y$:

- a key dependency in F if $Y = U$
- a primary-key dependency: only one key exists

Example

$\forall.\ \neg Par(x) \vee Ma(x) \vee Fa(x)$

Example

$\forall.\ \neg M(n, s, m) \vee \neg M(m, t, w) \vee s \leq t$

Example primary-key dependency

Name $\rightarrow$ Address Salary

Constraint classes

Universal constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n \vee B_1 \vee \cdots \vee B_m$

Denial constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n$

Functional dependencies

$X \rightarrow Y$:

- a key dependency in F if $Y = U$
- a primary-key dependency: only one key exists

Inclusion dependencies

$R[X] \subseteq S[Y]$:

- a foreign key constraint if Y is a key of S

Example

$\forall.\ \neg Par(x) \vee Ma(x) \vee Fa(x)$

Example

$\forall.\ \neg M(n, s, m) \vee \neg M(m, t, w) \vee s \leq t$

Example primary-key dependency

Name $\rightarrow$ Address Salary

Universal constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n \vee B_1 \vee \cdots \vee B_m$

Example

$\forall.\ \neg Par(x) \vee Ma(x) \vee Fa(x)$

Denial constraints

$\forall.\ \neg A_1 \vee \cdots \vee \neg A_n$

Example

$\forall.\ \neg M(n, s, m) \vee \neg M(m, t, w) \vee s \leq t$

Functional dependencies

$X \rightarrow Y$:

- a key dependency in F if $Y = U$
- a primary-key dependency: only one key exists

Example primary-key dependency

Name $\rightarrow$ Address Salary

Inclusion dependencies

$R[X] \subseteq S[Y]$:

- a foreign key constraint if Y is a key of S

Example foreign key constraint

$M[Manager] \subseteq M[Name]$

Building queries that compute CQAs

- relational calculus (algebra) $\leadsto$ relational calculus (algebra)
- SQL $\leadsto$ SQL
- leads to PTIME data complexity

Building queries that compute CQAs

- relational calculus (algebra) $\rightsquigarrow$ relational calculus (algebra)
- SQL $\rightsquigarrow$ SQL
- leads to PTIME data complexity

Query

$Emp(x, y, z)$

Query Rewriting

Building queries that compute CQAs

- relational calculus (algebra) $\rightsquigarrow$ relational calculus (algebra)
- SQL $\rightsquigarrow$ SQL
- leads to PTIME data complexity

Query

$Emp(x, y, z)$

Integrity constraint

$\forall\ x, y, z, y', z'.\ \neg Emp(x, y, z) \lor \neg Emp(x, y', z') \lor z = z'$

Query Rewriting

Building queries that compute CQAs

- relational calculus (algebra) $\rightsquigarrow$ relational calculus (algebra)
- SQL $\rightsquigarrow$ SQL
- leads to PTIME data complexity

Query

$Emp(x, y, z)$

Integrity constraint

$\forall\, x, y, z, y', z'.\ \neg Emp(x, y, z) \vee \neg Emp(x, y', z') \vee z = z'$

Query Rewriting

Building queries that compute CQAs

- relational calculus (algebra) $\rightsquigarrow$ relational calculus (algebra)
- SQL $\rightsquigarrow$ SQL
- leads to PTIME data complexity

Query

$Emp(x, y, z)$

Integrity constraint

$\forall\ x, y, z, y', z'.\ \neg Emp(x, y, z) \vee \neg Emp(x, y', z') \vee z = z'$

Rewritten query

$Emp(x, y, z) \wedge \forall\ y', z'.\ \neg Emp(x, y', z') \vee z = z'$

(Arenas, Bertossi, Ch. [ABC99])

- Integrity constraints: binary universal
- Queries: conjunctions of literals (relational algebra: $\sigma, \times, -$)

(Arenas, Bertossi, Ch. [ABC99])

- Integrity constraints: binary universal
- Queries: conjunctions of literals (relational algebra: $\sigma, \times, -$)

(Fuxman, Miller [FM07])

- Integrity constraints: primary key functional dependencies
- Queries: C_{forest}
  - a class of conjunctive queries ($\pi, \sigma, \times$)
  - no cycles
  - no non-key or non-full joins
  - no repeated relation symbols
  - no built-ins
- Generalization: conjunctive queries expressed as rooted rules (Wijsen [Wij07])

SQL query

```
SELECT Name FROM Emp
WHERE Salary ≥ 10K
```

SQL Rewriting

SQL query

```
SELECT Name FROM Emp
WHERE Salary ≥ 10K
```

SQL rewritten query

```
SELECT e1.Name FROM Emp e1
WHERE e1.Salary ≥ 10K AND NOT EXISTS
       (SELECT * FROM EMPLOYEE e2
       WHERE e2.Name = e1.Name AND e2.Salary < 10K)
```

SQL Rewriting

SQL query

```
SELECT Name FROM Emp
WHERE Salary ≥ 10K
```

SQL rewritten query

```
SELECT e1.Name FROM Emp e1
WHERE e1.Salary ≥ 10K AND NOT EXISTS
       (SELECT * FROM EMPLOYEE e2
       WHERE e2.Name = e1.Name AND e2.Salary < 10K)
```

(Fuxman, Fazli, Miller [FM05])

- ConQuer: a system for computing CQAs
- conjunctive (C_{forest}) and aggregation SQL queries
- databases can be annotated with consistency indicators
- tested on TPC-H queries and medium-size databases

Conflict Hypergraph

Vertices

Tuples in the database.

(Gates, Redmond, 20M)

(Grove, Santa Clara, 10M)

(Gates, Redmond, 30M)

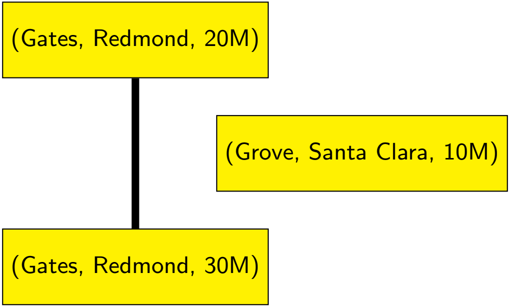
Conflict Hypergraph

Vertices

Tuples in the database.

Edges

Minimal sets of tuples violating a constraint.



Conflict Hypergraph

Vertices

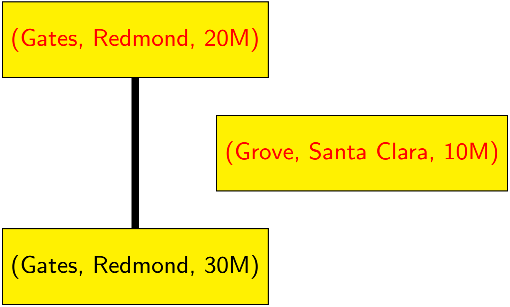
Tuples in the database.

Edges

Minimal sets of tuples violating a constraint.

Repairs

Maximal independent sets in the conflict graph.



Conflict Hypergraph

Vertices

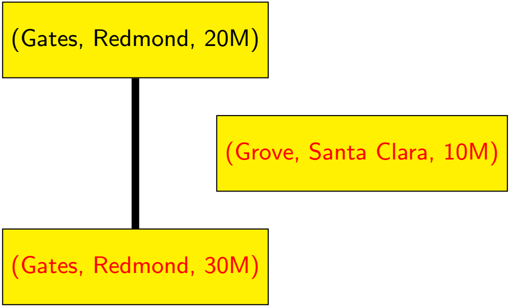
Tuples in the database.

Edges

Minimal sets of tuples violating a constraint.

Repairs

Maximal independent sets in the conflict graph.



Algorithm HProver

INPUT: query Φ a disjunction of ground literals, conflict hypergraph G
OUTPUT: is Φ false in some repair of D w.r.t. IC?
ALGORITHM:

1. $\neg\Phi = P_1(t_1) \wedge \cdots \wedge P_m(t_m) \wedge \neg P_{m+1}(t_{m+1}) \wedge \cdots \wedge \neg P_n(t_n)$
2. find a consistent set of facts S such that
   - $S \supseteq \{P_1(t_1), \ldots, P_m(t_m)\}$
   - for every fact $A \in \{P_{m+1}(t_{m+1}), \ldots, P_n(t_n)\}$: $A \notin D$ or there is an edge $E = \{A, B_1, \ldots, B_m\}$ in G and $S \supseteq \{B_1, \ldots, B_m\}$.

Algorithm HProver

INPUT: query Φ a disjunction of ground literals, conflict hypergraph G
OUTPUT: is Φ false in some repair of D w.r.t. IC?
ALGORITHM:

1. $\neg\Phi = P_1(t_1) \wedge \cdots \wedge P_m(t_m) \wedge \neg P_{m+1}(t_{m+1}) \wedge \cdots \wedge \neg P_n(t_n)$
2. find a consistent set of facts S such that
   - $S \supseteq \{P_1(t_1), \ldots, P_m(t_m)\}$
   - for every fact $A \in \{P_{m+1}(t_{m+1}), \ldots, P_n(t_n)\}$: $A \notin D$ or there is an edge $E = \{A, B_1, \ldots, B_m\}$ in G and $S \supseteq \{B_1, \ldots, B_m\}$.

(Ch., Marcinkowski, Staworko [CMS04])

- Hippo: a system for computing CQAs in PTIME
- quantifier-free queries and denial constraints
- only edges of the conflict hypergraph are kept in main memory
- optimization can eliminate many (sometimes all) database accesses in HProver
- tested for medium-size synthetic databases

Logic programs

Specifying repairs as answer sets of logic programs

- (Arenas, Bertossi, Ch. [ABC03])
- (Greco, Greco, Zumpano [GGZ03])
- (Calì, Lembo, Rosati [CLR03b])

Specifying repairs as answer sets of logic programs

- (Arenas, Bertossi, Ch. [ABC03])
- (Greco, Greco, Zumpano [GGZ03])
- (Calì, Lembo, Rosati [CLR03b])

Example

$emp(x, y, z) \leftarrow emp_D(x, y, z), not\ dubious\_emp(x, y, z).$
$dubious\_emp(x, y, z) \leftarrow emp_D(x, y, z), emp(x, y', z'), y \neq y'.$
$dubious\_emp(x, y, z) \leftarrow emp_D(x, y, z), emp(x, y', z'), z \neq z'.$

Specifying repairs as answer sets of logic programs

- (Arenas, Bertossi, Ch. [ABC03])
- (Greco, Greco, Zumpano [GGZ03])
- (Calì, Lembo, Rosati [CLR03b])

Example

$emp(x, y, z) \leftarrow emp_D(x, y, z), \textit{not}\ dubious\_emp(x, y, z).$
$dubious\_emp(x, y, z) \leftarrow emp_D(x, y, z), emp(x, y', z'), y \neq y'.$
$dubious\_emp(x, y, z) \leftarrow emp_D(x, y, z), emp(x, y', z'), z \neq z'.$

Answer sets

- $\{emp(Gates, Redmond, 20M), emp(Grove, SantaClara, 10M), \ldots\}$
- $\{emp(Gates, Redmond, 30M), emp(Grove, SantaClara, 10M), \ldots\}$

Logic Programs for computing CQAs

Logic Programs

- disjunction and classical negation
- checking whether an atom is in all answer sets is Π_2^p-complete
- `dlv`, `smodels`, ...

Logic Programs for computing CQAs

Logic Programs

- disjunction and classical negation
- checking whether an atom is in all answer sets is Π_2^p-complete
- `dlv`, `smodels`, ...

Scope

- arbitrary first-order queries
- universal constraints
- approach unlikely to yield tractable cases

Logic Programs for computing CQAs

Logic Programs

- disjunction and classical negation
- checking whether an atom is in all answer sets is Π_2^p-complete
- `dlv`, `smodels`, ...

Scope

- arbitrary first-order queries
- universal constraints
- approach unlikely to yield tractable cases

INFOMIX (Eiter et al. [EFGL03])

- combines CQA with data integration (GAV)
- uses `dlv` for repair computations
- optimization techniques: localization, factorization
- tested on small-to-medium-size legacy databases

Inconsistency and Incompleteness

There is a clear correspondence between repairs and possible worlds.

There is a clear correspondence between repairs and possible worlds.

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

There is a clear correspondence between repairs and possible worlds.

Name	City	Salary
Gates	Redmond	20M
Gates	Redmond	30M
Jobs	Cupertino	10M

Name $\rightarrow$ City Salary

$$Emp(Gates, Redmond, 20M) \vee Emp(Gates, Redmond, 30M)$$

$$Emp(Jobs, Cupertino, 10M)$$

Minimal models

Minimal models

Repairs vs. minimal models (Molinaro, Ch., Marcinkowski, 2008)

- For denial constraints, the set of repairs of an instance can be represented as the set of minimal models of a disjunctive database of at most exponential size
- The size of the disjunctive database is polynomial for a primary-key FD but may be exponential for two key FDs or one non-key FD.

Minimal models

Repairs vs. minimal models (Molinaro, Ch., Marcinkowski, 2008)

- For denial constraints, the set of repairs of an instance can be represented as the set of minimal models of a disjunctive database of at most exponential size
- The size of the disjunctive database is polynomial for a primary-key FD but may be exponential for two key FDs or one non-key FD.

Conclusion

It is unlikely that new tractable cases of CQA will be obtained through the disjunctive database representation of the repairs:

Repairs vs. minimal models (Molinaro, Ch., Marcinkowski, 2008)

- For denial constraints, the set of repairs of an instance can be represented as the set of minimal models of a disjunctive database of at most exponential size
- The size of the disjunctive database is polynomial for a primary-key FD but may be exponential for two key FDs or one non-key FD.

Conclusion

It is unlikely that new tractable cases of CQA will be obtained through the disjunctive database representation of the repairs:

- but (Imielinski, van der Meyden, Vadaparty [IvdMV95])...

Name	City	Salary
Gates	Redmond	OR(20M,30M)
Jobs	Cupertino	10M

Theorem (Ch., Marcinkowski [CM05a])

For primary-key functional dependencies and conjunctive queries, consistent query answering is data-complete for co-NP.

Co-NP-completeness of CQA

Theorem (Ch., Marcinkowski [CM05a])

For primary-key functional dependencies and conjunctive queries, consistent query answering is data-complete for co-NP.

Proof.

Membership: S is a repair iff $S \models IC$ and $W \not\models IC$ if $W = S \cup M$.
Co-NP-hardness: reduction from MONOTONE 3-SAT.

1. Positive clauses $\beta_1 = \phi_1 \wedge \cdots \wedge \phi_m$, negative clauses $\beta_2 = \psi_{m+1} \wedge \cdots \wedge \psi_l$.
2. Database D contains two binary relations $R(A, B)$ and $S(A, B)$:
   - $R(i, p)$ if variable p occurs in ϕ_i, $i = 1, \ldots, m$.
   - $S(i, p)$ if variable p occurs in ψ_i, $i = m + 1, \ldots, l$.
3. A is the primary key of both R and S.
4. Query $Q \equiv \exists x, y, z.\ \big(R(x, y) \wedge S(z, y)\big)$.
5. There is an assignment which satisfies $\beta_1 \wedge \beta_2$ iff there exists a repair in which Q is false.

$\square$

Co-NP-completeness of CQA

Theorem (Ch., Marcinkowski [CM05a])

For primary-key functional dependencies and conjunctive queries, consistent query answering is data-complete for co-NP.

Proof.

Membership: S is a repair iff $S \models IC$ and $W \not\models IC$ if $W = S \cup M$.
Co-NP-hardness: reduction from MONOTONE 3-SAT.

1. Positive clauses $\beta_1 = \phi_1 \wedge \cdots \wedge \phi_m$, negative clauses $\beta_2 = \psi_{m+1} \wedge \cdots \wedge \psi_l$.
2. Database D contains two binary relations $R(A, B)$ and $S(A, B)$:
   - $R(i, p)$ if variable p occurs in ϕ_i, $i = 1, \ldots, m$.
   - $S(i, p)$ if variable p occurs in ψ_i, $i = m + 1, \ldots, l$.
3. A is the primary key of both R and S.
4. Query $Q \equiv \exists x, y, z. \, \big(R(x, y) \wedge S(z, y)\big)$.
5. There is an assignment which satisfies $\beta_1 \wedge \beta_2$ iff there exists a repair in which Q is false.

□

Q does not belong to C_{forest}.

Data complexity of CQA

	Primary keys	*Arbitrary keys*	*Denial*	*Universal*
$\sigma, \times, -$				
$\sigma, \times, -, \cup$				
σ, π				
$\sigma, \pi, \times$				
$\sigma, \pi, \times, -, \cup$				

Data complexity of CQA

	Primary keys	*Arbitrary keys*	*Denial*	*Universal*
$\sigma, \times, -$	PTIME	PTIME		PTIME: binary
$\sigma, \times, -, \cup$				
σ, π				
$\sigma, \pi, \times$				
$\sigma, \pi, \times, -, \cup$				

- (Arenas, Bertossi, Ch. [ABC99])

Data complexity of CQA

	Primary keys	*Arbitrary keys*	*Denial*	*Universal*
$\sigma, \times, -$	PTIME	PTIME	PTIME	PTIME: binary
$\sigma, \times, -, \cup$	PTIME	PTIME	PTIME	
σ, π	PTIME	co-NPC	co-NPC	
$\sigma, \pi, \times$	co-NPC	co-NPC	co-NPC	
$\sigma, \pi, \times, -, \cup$	co-NPC	co-NPC	co-NPC	

- (Arenas, Bertossi, Ch. [ABC99])
- (Ch., Marcinkowski [CM05a])

Data complexity of CQA

	Primary keys	*Arbitrary keys*	*Denial*	*Universal*
$\sigma, \times, -$	PTIME	PTIME	PTIME	PTIME: binary
$\sigma, \times, -, \cup$	PTIME	PTIME	PTIME	
σ, π	PTIME	co-NPC	co-NPC	
$\sigma, \pi, \times$	co-NPC PTIME: C_{forest}	co-NPC	co-NPC	
$\sigma, \pi, \times, -, \cup$	co-NPC	co-NPC	co-NPC	

- (Arenas, Bertossi, Ch. [ABC99])
- (Ch., Marcinkowski [CM05a])
- (Fuxman, Miller [FM07])

Data complexity of CQA

	Primary keys	*Arbitrary keys*	*Denial*	*Universal*
$\sigma, \times, -$	PTIME	PTIME	PTIME	PTIME: binary Π_2^p-complete
$\sigma, \times, -, \cup$	PTIME	PTIME	PTIME	Π_2^p-complete
σ, π	PTIME	co-NPC	co-NPC	Π_2^p-complete
$\sigma, \pi, \times$	co-NPC PTIME: C_{forest}	co-NPC	co-NPC	Π_2^p-complete
$\sigma, \pi, \times, -, \cup$	co-NPC	co-NPC	co-NPC	Π_2^p-complete

- (Arenas, Bertossi, Ch. [ABC99])
- (Ch., Marcinkowski [CM05a])
- (Fuxman, Miller [FM07])
- (Staworko, Ph.D., 2007), (Staworko, Ch., 2008):
  - quantifier-free queries
  - co-NPC for full TGDs and denial constraints
  - PTIME for acyclic full TGDs, join dependencies and denial constraints

The Explosion of Semantics

Tuple-based repairs

- asymmetric treatment of insertion and deletion:
  - repairs by minimal deletions only (Ch., Marcinkowski [CM05a]): data possibly incorrect but complete
  - repairs by minimal deletions and arbitrary insertions (Calì, Lembo, Rosati [CLR03a]): data possibly incorrect and incomplete
- minimal cardinality changes (Lopatenko, Bertossi [LB07])
- more in the pipeline...

The Explosion of Semantics

Tuple-based repairs

- asymmetric treatment of insertion and deletion:
  - repairs by minimal deletions only (Ch., Marcinkowski [CM05a]): data possibly incorrect but complete
  - repairs by minimal deletions and arbitrary insertions (Calì, Lembo, Rosati [CLR03a]): data possibly incorrect and incomplete
- minimal cardinality changes (Lopatenko, Bertossi [LB07])
- more in the pipeline...

Attribute-based repairs

- (A) ground and non-ground repairs (Wijsen [Wij05])
- (B) project-join repairs (Wijsen [Wij06])
- (C) repairs minimizing Euclidean distance (Bertossi et al. [BBFL08])
- (D) repairs of minimum cost (Bohannon et al. [BFFR05])

The Explosion of Semantics

Tuple-based repairs

- asymmetric treatment of insertion and deletion:
  - repairs by minimal deletions only (Ch., Marcinkowski [CM05a]): data possibly incorrect but complete
  - repairs by minimal deletions and arbitrary insertions (Calì, Lembo, Rosati [CLR03a]): data possibly incorrect and incomplete
- minimal cardinality changes (Lopatenko, Bertossi [LB07])
- more in the pipeline...

Attribute-based repairs

- (A) ground and non-ground repairs (Wijsen [Wij05])
- (B) project-join repairs (Wijsen [Wij06])
- (C) repairs minimizing Euclidean distance (Bertossi et al. [BBFL08])
- (D) repairs of minimum cost (Bohannon et al. [BFFR05])

Computational complexity

- (A) and (B): similar to tuple based repairs
- (C) and (D): checking existence of a repair of cost $< K$ NP-complete.

The Need for Attribute-based Repairing

Tuple-based repairing leads to information loss.

Tuple-based repairing leads to information loss.

EmpDept

Name	*Dept*	*Location*
John	Sales	Buffalo
Mary	Sales	Toronto

Name $\longrightarrow$ Dept

Dept $\longrightarrow$ City

The Need for Attribute-based Repairing

Tuple-based repairing leads to information loss.

EmpDept

Name	*Dept*	*Location*
John	Sales	Buffalo
Mary	Sales	Toronto

Name → Dept

Dept → City

Name	*Dept*	*Location*
John	Sales	Buffalo

Name → Dept

Dept → City

Name	*Dept*	*Location*
Mary	Sales	Toronto

Name → Dept

Dept → City

Attribute-based Repairs through Tuple-based Repairs (Wijsen [Wij06])

Repair the lossless join decomposition:

$$\pi_{Name,Dept}(EmpDept) \bowtie \pi_{Dept,Location}(EmpDept)$$

Repair the lossless join decomposition:

$$\pi_{Name,Dept}(EmpDept) \bowtie \pi_{Dept,Location}(EmpDept)$$

Name	*Dept*	*Location*
John	Sales	Buffalo
John	Sales	Toronto
Mary	Sales	Buffalo
Mary	Sales	Toronto

Name $\rightarrow$ Dept

Dept $\rightarrow$ City

Repair the lossless join decomposition:

$$\pi_{Name,Dept}(EmpDept) \bowtie \pi_{Dept,Location}(EmpDept)$$

Name	*Dept*	*Location*
John	Sales	Buffalo
John	Sales	Toronto
Mary	Sales	Buffalo
Mary	Sales	Toronto

Name $\rightarrow$ Dept

Dept $\rightarrow$ City

Name	*Dept*	*Location*
John	Sales	Buffalo
Mary	Sales	Buffalo

Name $\rightarrow$ Dept

Dept $\rightarrow$ City

Name	*Dept*	*Location*
John	Sales	Toronto
Mary	Sales	Toronto

Name $\rightarrow$ Dept

Dept $\rightarrow$ City

Probabilistic framework for "dirty" databases

(Andritsos, Fuxman, Miller [AFM06])

- potential duplicates identified and grouped into clusters
- worlds $\approx$ repairs: one tuple from each cluster
- world probability: product of tuple probabilities
- clean answers: in the query result in some (supporting) world
- clean answer probability: sum of the probabilities of supporting worlds
  - consistent answer: clean answer with probability 1

Probabilistic framework for "dirty" databases

(Andritsos, Fuxman, Miller [AFM06])

- potential duplicates identified and grouped into clusters
- worlds $\approx$ repairs: one tuple from each cluster
- world probability: product of tuple probabilities
- clean answers: in the query result in some (supporting) world
- clean answer probability: sum of the probabilities of supporting worlds
  - consistent answer: clean answer with probability 1

Salaries with probabilities

EmpProb

Name	*Salary*	Prob
Gates	20M	0.7
Gates	30M	0.3
Jobs	10M	0.5
Jobs	20M	0.5

SQL query

```
SELECT Name
FROM EmpProb e
WHERE e.Salary > 15M
```

Computing Clean Answers

SQL query

```
SELECT Name
FROM EmpProb e
WHERE e.Salary > 15M
```

SQL rewritten query

```
SELECT e.Name,SUM(e.Prob)
FROM EmpProb e
WHERE e.Salary > 15M
GROUP BY e.Name
```

SQL query

```
SELECT Name
FROM EmpProb e
WHERE e.Salary > 15M
```

SQL rewritten query

```
SELECT e.Name,SUM(e.Prob)
FROM EmpProb e
WHERE e.Salary > 15M
GROUP BY e.Name
```

EmpProb

Name	*Salary*	Prob
Gates	20M	0.7
Gates	30M	0.3
Jobs	10M	0.5
Jobs	20M	0.5

Name $\rightarrow$ Salary

Computing Clean Answers

SQL query

```
SELECT Name
FROM EmpProb e
WHERE e.Salary > 15M
```

SQL rewritten query

```
SELECT e.Name,SUM(e.Prob)
FROM EmpProb e
WHERE e.Salary > 15M
GROUP BY e.Name
```

EmpProb

Name	*Salary*	Prob
Gates	20M	0.7
Gates	30M	0.3
Jobs	10M	0.5
Jobs	20M	0.5

Name → Salary

```
SELECT e.Name,SUM(e.Prob)
FROM EmpProb e
WHERE e.Salary > 15M
GROUP BY e.Name
```

SQL query

```
SELECT Name
FROM EmpProb e
WHERE e.Salary > 15M
```

SQL rewritten query

```
SELECT e.Name,SUM(e.Prob)
FROM EmpProb e
WHERE e.Salary > 15M
GROUP BY e.Name
```

EmpProb

Name	*Salary*	Prob
Gates	20M	0.7
Gates	30M	0.3
Jobs	10M	0.5
Jobs	20M	0.5

Name $\rightarrow$ Salary

```
SELECT e.Name,SUM(e.Prob)
FROM EmpProb e
WHERE e.Salary > 15M
GROUP BY e.Name
```

$\longrightarrow$

Name	Prob
Gates	1
Jobs	0.5

Taking Stock: Good News

Technology

- practical methods for CQA for subsets of SQL:
  - restricted conjunctive/aggregation queries, primary/foreign-key constraints
  - quantifier-free queries, denial constraints/acyclic TGDs/JDs
  - LP-based approaches for expressive query/constraint languages
- (slow) emergence of generic techniques
- implemented in prototype systems
- tested on medium-size databases

Taking Stock: Good News

Technology

- practical methods for CQA for subsets of SQL:
  - restricted conjunctive/aggregation queries, primary/foreign-key constraints
  - quantifier-free queries, denial constraints/acyclic TGDs/JDs
  - LP-based approaches for expressive query/constraint languages
- (slow) emergence of generic techniques
- implemented in prototype systems
- tested on medium-size databases

The CQA Community

- over 30 active researchers
- over 100 publications (since 1999)
- at least 8 doctoral dissertations in Europe and North America
- 2007 SIGMOD Doctoral Dissertation Award (Ariel Fuxman)
- overview papers [BC03, Ber06, Cho07, CM05b]

Taking Stock: Initial Progress

Taking Stock: Initial Progress

"Blending in" CQA

- data integration: tension between repairing and satisfying source-to-target dependencies

Taking Stock: Initial Progress

"Blending in" CQA

- data integration: tension between repairing and satisfying source-to-target dependencies

Extensions

- uncertainty:
  - inconsistency leads to but cannot be reduced to uncertainty
  - repairs vs. possible worlds
  - probabilistic data
  - nulls: SQL conformance
- priorities:
  - preferred repairs
  - application: conflict resolution
- XML
  - notions of integrity constraint and repair
  - repair minimality based on tree edit distance?
- aggregate constraints

Taking Stock: Largely Open Issues

Taking Stock: Largely Open Issues

Applications

- no deployed applications
- repairing vs. CQA: data and query characteristics
- heuristics for CQA and repairing

Taking Stock: Largely Open Issues

Applications

- no deployed applications
- repairing vs. CQA: data and query characteristics
- heuristics for CQA and repairing

CQA in context

- taming the semantic explosion
- CQA and data cleaning
- CQA and schema matching/mapping

Taking Stock: Largely Open Issues

Applications

- no deployed applications
- repairing vs. CQA: data and query characteristics
- heuristics for CQA and repairing

CQA in context

- taming the semantic explosion
- CQA and data cleaning
- CQA and schema matching/mapping

Foundations

- repair checking
- defining measures of consistency
- proving non-existence of rewritings

Taking Stock: Largely Open Issues

Applications

- no deployed applications
- repairing vs. CQA: data and query characteristics
- heuristics for CQA and repairing

CQA in context

- taming the semantic explosion
- CQA and data cleaning
- CQA and schema matching/mapping

Foundations

- repair checking
- defining measures of consistency
- proving non-existence of rewritings

Acknowledgments

Marcelo Arenas
Alessandro Artale
Leo Bertossi
Loreto Bravo
Andrea Calì
Thomas Eiter
Wenfei Fan
Enrico Franconi
Ariel Fuxman
Gianluigi Greco
Sergio Greco
Phokion Kolaitis
Domenico Lembo
Maurizio Lenzerini
Jerzy Marcinkowski
Renée Miller
Cristian Molinaro
Riccardo Rosati
Sławek Staworko
David Toman
Jef Wijsen

M. Arenas, L. Bertossi, and J. Chomicki.
Consistent Query Answers in Inconsistent Databases.
In *ACM Symposium on Principles of Database Systems (PODS)*, pages 68–79, 1999.

M. Arenas, L. Bertossi, and J. Chomicki.
Answer Sets for Consistent Query Answering in Inconsistent Databases.
Theory and Practice of Logic Programming, 3(4–5):393–424, 2003.

P. Andritsos, A. Fuxman, and R. Miller.
Clean Answers over Dirty Databases.
In *IEEE International Conference on Data Engineering (ICDE)*, 2006.

L. Bertossi, L. Bravo, E. Franconi, and A. Lopatenko.
The Complexity and Approximation of Fixing Numerical Attributes in Databases under Integrity Constraints.
Information Systems, 33(4-5):407–434, June-July 2008.

L. Bertossi and J. Chomicki.
Query Answering in Inconsistent Databases.
In J. Chomicki, R. van der Meyden, and G. Saake, editors, *Logics for Emerging Applications of Databases*, pages 43–83. Springer-Verlag, 2003.

L. Bertossi.
Consistent Query Answering in Databases.
SIGMOD Record, 35(2), June 2006.

P. Bohannon, M. Flaster, W. Fan, and R. Rastogi.
A Cost-Based Model and Effective Heuristic for Repairing Constraints by Value Modification.
In *ACM SIGMOD International Conference on Management of Data*, pages 143–154, 2005.

A. Borgida.
Language Features for Flexible Handling of Exceptions in Information Systems.
ACM Transactions on Database Systems, 10(4):565–603, 1985.

J. Chomicki.
Consistent Query Answering: Five Easy Pieces.
In *International Conference on Database Theory (ICDT)*, pages 1–17. Springer, LNCS 4353, 2007.
Keynote talk.

A. Calì, D. Lembo, and R. Rosati.
On the Decidability and Complexity of Query Answering over Inconsistent and Incomplete Databases.
In *ACM Symposium on Principles of Database Systems (PODS)*, pages 260–271, 2003.

A. Calì, D. Lembo, and R. Rosati.
Query Rewriting and Answering under Constraints in Data Integration Systems.

In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 16–21, 2003.

J. Chomicki and J. Marcinkowski.
Minimal-Change Integrity Maintenance Using Tuple Deletions.
Information and Computation, 197(1-2):90–121, 2005.

J. Chomicki and J. Marcinkowski.
On the Computational Complexity of Minimal-Change Integrity Maintenance in Relational Databases.
In L. Bertossi, A. Hunter, and T. Schaub, editors, *Inconsistency Tolerance*, pages 119–150. Springer-Verlag, 2005.

J. Chomicki, J. Marcinkowski, and S. Staworko.
Computing Consistent Query Answers Using Conflict Hypergraphs.
In *International Conference on Information and Knowledge Management (CIKM)*, pages 417–426. ACM Press, 2004.

P. De Bra and J. Paredaens.
Conditional Dependencies for Horizontal Decompositions.
In *International Colloquium on Automata, Languages and Programming (ICALP)*, pages 123–141, 1983.

T. Eiter, M. Fink, G. Greco, and D. Lembo.
Efficient Evaluation of Logic Programs for Querying Data Integration Systems.
In *International Conference on Logic Programming (ICLP)*, pages 163–177, 2003.

A. Fuxman and R. J. Miller.
ConQuer: Efficient Management of Inconsistent Databases.
In *ACM SIGMOD International Conference on Management of Data*, pages 155–166, 2005.

A. Fuxman and R. J. Miller.
First-Order Query Rewriting for Inconsistent Databases.
Journal of Computer and System Sciences, 73(4):610–635, 2007.

G. Greco, S. Greco, and E. Zumpano.
A Logical Framework for Querying and Repairing Inconsistent Databases.
IEEE Transactions on Knowledge and Data Engineering, 15(6):1389–1408, 2003.

T. Imieliński, R. van der Meyden, and K. Vadaparty.
Complexity Tailored Design: A New Design Methodology for Databases With Incomplete Information.
Journal of Computer and System Sciences, 51(3):405–432, 1995.

A. Lopatenko and L. Bertossi.
Complexity of Consistent Query Answering in Databases under Cardinality-Based and Incremental Repair Semantics.
In *International Conference on Database Theory (ICDT)*, pages 179–193. Springer, LNCS 4353, 2007.

J. Wijsen.
Database Repairing Using Updates.

ACM Transactions on Database Systems, 30(3):722–768, 2005.

J. Wijsen.
Project-Join Repair: An Approach to Consistent Query Answering Under Functional Dependencies.
In *International Conference on Flexible Query Answering Systems (FQAS)*, 2006.

J. Wijsen.
On the Consistent Rewriting of Conjunctive Queries Under Primary Key Constraints.
In *International Workshop on Database Programming Languages*, pages 112–126. Springer, LNCS 4797, 2007.