Incomplete Databases

A Complete Database

Flight	origin	destination	airline
	VIE	LHR	BA
	LHR	EDI	BA
	LGW	GLA	U2
	EDI	VIE	OS

Airport	code	city
	VIE	Vienna
	LHR	London
	LGW	London
	LCA	Larnaca
	GLA	Glasgow
	EDI	Edinburgh

An Incomplete Database

Flight	origin	destination	airline
	VIE	LHR	$\bot_1$
	LHR	EDI	$\bot_1$
	$\bot_4$	GLA	U2
	EDI	VIE	OS

Airport	code	city
	VIE	Vienna
	LHR	$\bot_2$
	LGW	London
	LCA	$\bot_3$
	GLA	Glasgow
	EDI	Edinburgh

$\bot_1$, $\bot_2$, $\bot_3$, $\bot_4$ - marked (also called labeled) nulls

An Incomplete Database

Flight	origin	destination	airline
	VIE	LHR	$\bot_1$
	LHR	EDI	$\bot_1$
	$\bot_4$	GLA	U2
	EDI	VIE	OS

Airport	code	city
	VIE	Vienna
	LHR	$\bot_2$
	LGW	London
	LCA	$\bot_3$
	GLA	Glasgow
	EDI	Edinburgh

values are drawn from two countably disjoint infinite sets of values - **Const** and **Nulls**

Querying Incomplete Databases

Flight	origin	destination	airline
	VIE	LHR	$\perp_1$
	LHR	EDI	$\perp_1$
	$\perp_4$	GLA	U2
	EDI	VIE	OS

Airport	code	city
	VIE	Vienna
	LHR	$\perp_2$
	LGW	London
	LCA	$\perp_3$
	GLA	Glasgow
	EDI	Edinburgh

YES!

Q :- Airport(x,Vienna), Airport(y,London), Flight(x,y,z), Airport(w,Edinburgh), Flight(y,w,z)

Querying Incomplete Databases

Flight	origin	destination	airline
	VIE	LHR	$\perp_1$
	LHR	EDI	$\perp_1$
	$\perp_4$	GLA	U2
	EDI	VIE	OS

Airport	code	city
	VIE	Vienna
	LHR	$\perp_2$
	LGW	London
	LCA	$\perp_3$
	GLA	Glasgow
	EDI	Edinburgh

NO!

Q :- Airport(x,Vienna), Airport(y,London), Flight(x,y,BA), Airport(w,Edinburgh), Flight(y,w,z)

A Possible Completion - Closed World

Flight	origin	destination	airline
	VIE	LHR	BA
	LHR	EDI	BA
	LGW	GLA	U2
	EDI	VIE	OS

Airport	code	city
	VIE	Vienna
	LHR	London
	LGW	London
	LCA	Larnaca
	GLA	Glasgow
	EDI	Edinburgh

$\bot_1 \mapsto$ BA $\bot_2 \mapsto$ London $\bot_3 \mapsto$ Larnaca $\bot_4 \mapsto$ LGW

A Possible Completion - Closed World

Flight	origin	destination	airline
	VIE	LHR	AF
	LHR	EDI	AF
	LGW	GLA	U2
	EDI	VIE	OS

Airport	code	city
	VIE	Vienna
	LHR	London
	LGW	London
	LCA	Larnaca
	GLA	Glasgow
	EDI	Edinburgh

$\bot_1 \mapsto$ AF $\bot_2 \mapsto$ London $\bot_3 \mapsto$ Larnaca $\bot_4 \mapsto$ LGW

A Possible Completion - Open World

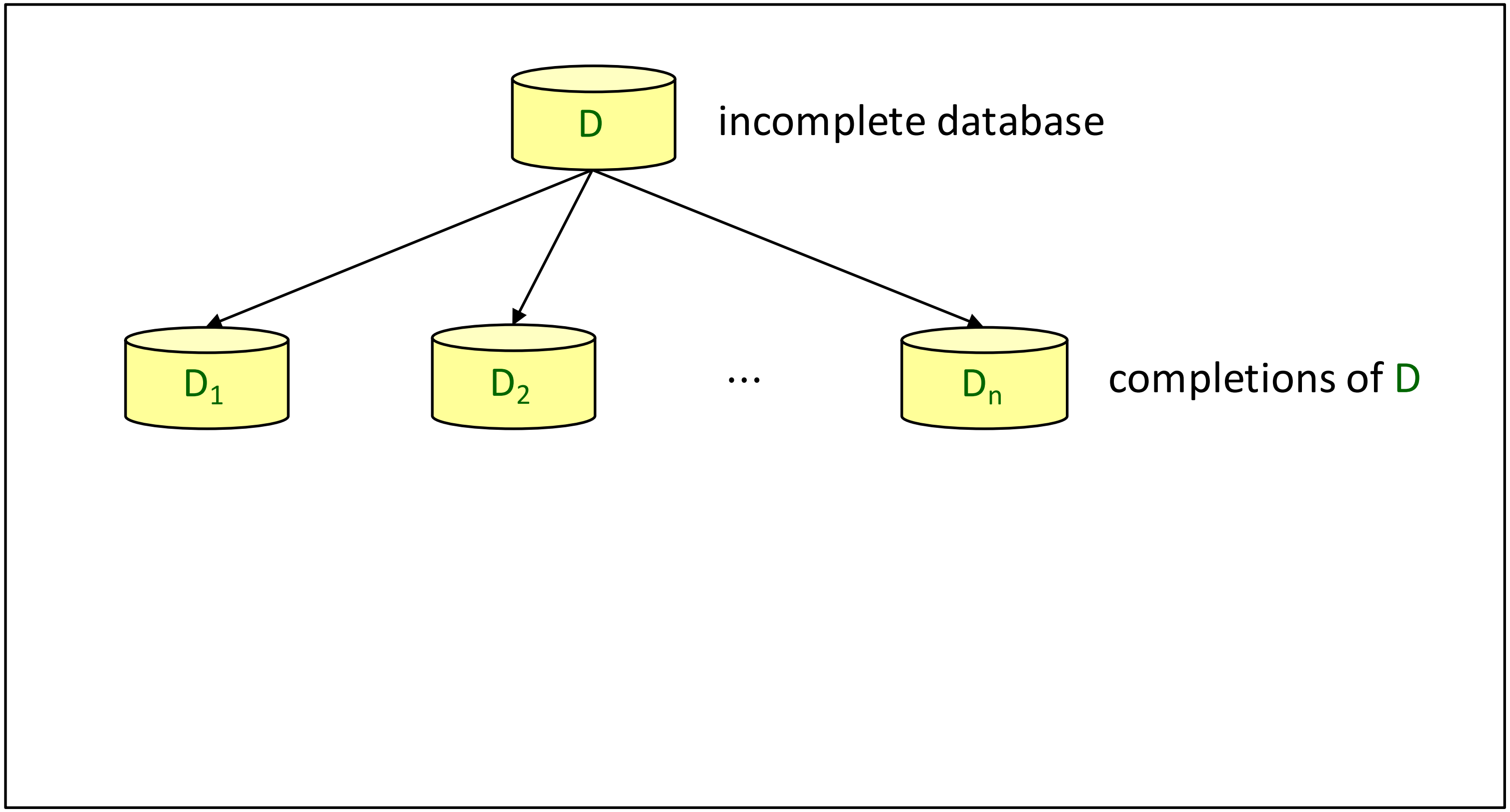
Flight	origin	destination	airline
	VIE	LHR	AF
	LHR	EDI	AF
	LGW	GLA	U2
	EDI	VIE	OS
	EDI	LGW	BE
	CDG	LHR	AF

Airport	code	city
	VIE	Vienna
	LHR	London
	LGW	London
	LCA	Larnaca
	GLA	Glasgow
	EDI	Edinburgh
	CDG	Paris

$\bot_1 \mapsto$ AF $\quad \bot_2 \mapsto$ London $\quad \bot_3 \mapsto$ Larnaca $\quad \bot_4 \mapsto$ LGW

Querying Incomplete Databases

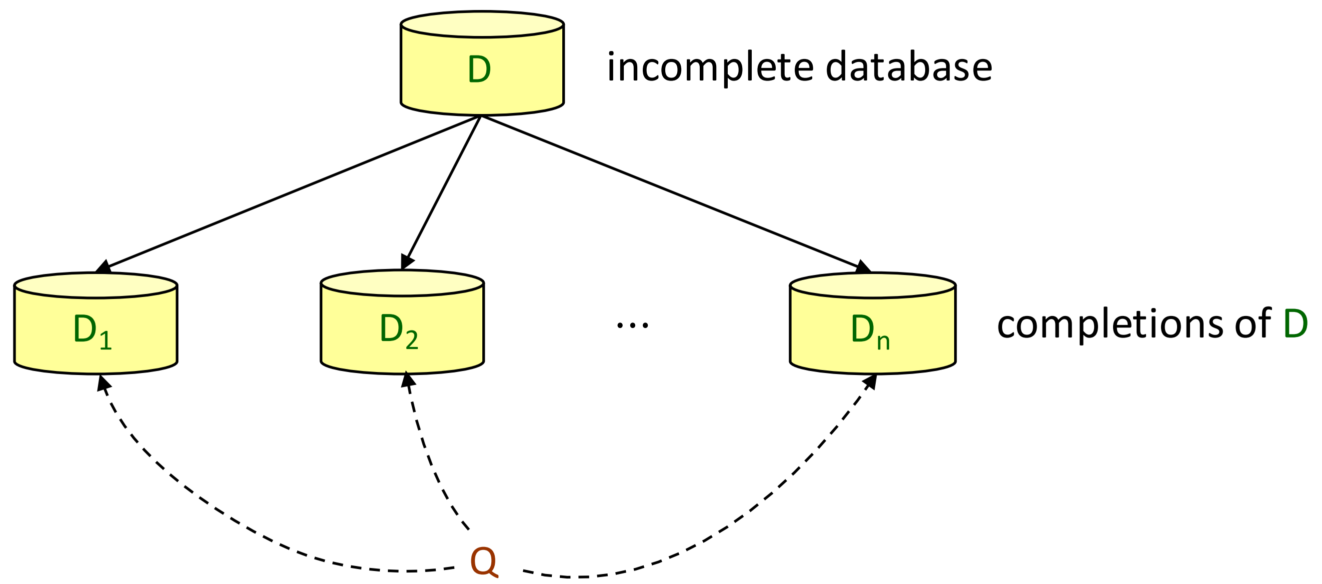
Closed/Open World Completion - a complete version of D



Querying Incomplete Databases

Closed/Open World Completion - a complete version of D

Certain answers - answers that are true in all completions



Answer(Q,D) = $Q(D_1) \cap \cdots \cap Q(D_n)$

Closed World Semantics

defined via valuations of nulls

the set of nulls occurring in D

A **valuation** of nulls on an incomplete database D is a function v : Null(D) → **Const**

CWA(D) = {v(D) | v is a valuation of nulls on D}

the database obtained from D after replacing the nulls according to v

Open World Semantics

defined via valuations of nulls

the set of nulls occurring in D

A **valuation** of nulls on an incomplete database D is a function $v : \text{Null}(D) \rightarrow \textbf{Const}$

$$\text{OWA}(D) = \{v(D) \cup D_0 \mid v \text{ is a valuation of nulls on } D\text{, and } D_0 \text{ is a complete database}\}$$

the database obtained from D after replacing the nulls according to v

Certain Answers

$$\text{Answer-CWA}(Q,D) = \bigcap_{C \in \text{CWA}(D)} Q(C)$$

$$\text{Answer-OWA}(Q,D) = \bigcap_{C \in \text{OWA}(D)} Q(C)$$

Note: tuples in certain answers cannot contain nulls, i.e., answer tuples consists only of constants

Querying Incomplete Databases

fix the semantics S ∈ {CWA, OWA}

IQA-S(**L**)

Input: an incomplete database D, a query Q/k ∈ **L**, a tuple of constants **t** ∈ $\mathbf{Const}^k$

Question: **t** ∈ Answer-S(Q,D)?

BIQA-S(**L**)

Input: an incomplete database D, a Boolean query Q ∈ **L**

Question: is Answer-S(Q,D) non-empty?

Theorem: IQA-S(**L**) $\equiv_L$ BIQA-S(**L**), where **L** ∈ {**RA**, **DRC**, **TRC, CQ**}

($\equiv_L$ means logspace-equivalent)

Complexity of BIQA-OWA

Theorem: For **L** $\in$ {**RA**, **DRC**, **TRC**}, BIQA-OWA(**L**) is undecidable

Proof: reduction from the validity problem for **L**

Validity

A Boolean query Q is **valid** if, for every database D, Q(D) is non-empty

VALID(**L**)

Input: a Boolean query Q ∈ **L**

Question: is Q valid?

Theorem: For **L** ∈ {**RA**, **DRC**, **TRC**}, VALID(**L**) is undecidable

Complexity of BIQA-OWA

Theorem: For **L** ∈ {**RA**, **DRC**, **TRC**}, BIQA-OWA(**L**) is undecidable

Proof: reduction from the validity problem for **L**

Let $D_\emptyset$ be the empty database

Answer-S(Q,$D_\emptyset$) is non-empty ⇔ for each $C \in$ OWA($D_\emptyset$), Q(C) is non-empty

but, OWA($D_\emptyset$) consists of all the databases

⇓

Answer-S(Q,$D_\emptyset$) is non-empty ⇔ Q is valid

Data Complexity of BIQA-S

input D, fixed Q

BIQA-S[Q](**L**)

Input: a database D

Question: is Answer-S(Q,D) non-empty?

Data Complexity of BIQA-CWA

Theorem: For **L** ∈ {**RA**, **DRC**, **TRC**}, the following hold:

- For every query Q ∈ **L**, BIQA-CWA[Q](**L**) is in coNP
- There exists a query Q ∈ **L** such that BIQA-CWA[Q](**L**) is coNP-hard

Proof:

- Guess a valuation of nulls on D, and check whether Q(v(D)) is empty
- Reduction from 3-Colorability to the complement of BIQA-CWA

3-Colorability

3COL

Input: an undirected graph **G** = (V,E)

Question: is there a function $c : V \to \{R,G,B\}$ such that $(v,u) \in E \Rightarrow c(v) \neq c(u)$?

coNP-hardness

Given an undirected graph **G** = (V,E)

construct a database D such that, for some fixed Q, it holds that

G is 3-colorable iff Answer-CWA(Q,D) is empty

$D = \{Node(\bot_u) : u \in V\} \cup \{Edge(\bot_u, \bot_v) : (u,v) \in E\}$

$$Q = \exists x \exists y \exists z \exists w\, (Node(x) \wedge Node(y) \wedge Node(z) \wedge Node(w) \wedge x \neq y \wedge x \neq z \wedge x \neq w \wedge y \neq z \wedge y \neq w \wedge z \neq w) \vee \exists x\, Edge(x,x)$$

Lemma: G is 3-colorable iff there is a valuation v of nulls on D such that Q(v(D)) is empty

Data Complexity of BIQA-CWA

Theorem: For **L** ∈ {**RA**, **DRC**, **TRC**}, the following hold:

- For every query Q ∈ **L**, BIQA-CWA[Q](**L**) is in coNP
- There exists a query Q ∈ **L** such that BIQA-CWA[Q](**L**) is coNP-hard

Proof:

- Guess a valuation of nulls on D, and check whether Q(v(D)) is empty
- Reduction from 3-Colorability to the complement of BIQA-CWA

but, what about conjunctive queries?

Naïve Evaluation

simply use the standard evaluation algorithm for complete databases

Theorem: Consider an incomplete database D, and a Boolean query Q ∈ **CQ**. Then

Q(D) = Answer-CWA(Q,D) and Q(D) = Answer-OWA(Q,D)

Proof:

(⇐) CWA(D) and OWA(D) contain a "copy" of D - replace all nulls by new constants. Therefore, if Answer-CWA(Q,D) or Answer-OWA(Q,D) is non-empty, then Q(D) is non-empty.

Naïve Evaluation

simply use the standard evaluation algorithm for complete databases

Theorem: Consider an incomplete database D, and a Boolean query Q ∈ **CQ**. Then

Q(D) = Answer-CWA(Q,D) and Q(D) = Answer-OWA(Q,D)

Proof:

(⇒) Q(D) is non-empty ⇒ Q → D

⇒ for every C, D → C implies Q → C

⇒ for every C, D → C implies Q(C) is non-empty

⇒ for every C ∈ CWA(D) ∪ OWA(D), Q(C) is non-empty

⇒ Answer-CWA(Q,D) and Answer-CWA(Q,D) are non-empty

Naïve Evaluation

simply use the standard evaluation algorithm for complete databases

Theorem: Consider an incomplete database D, and a Boolean query Q ∈ **CQ**. Then

Q(D) = Answer-CWA(Q,D) and Q(D) = Answer-OWA(Q,D)

Proof:

(⇒) Q(D) is non-empty ⇒ Q → D

⇒ **for every C, D → C implies Q → C**

⇒ for every C, D → C implies Q(C) is non-empty

⇒ for every C ∈ CWA(D) ∪ OWA(D), Q(C) is non-empty

⇒ Answer-CWA(Q,D) and Answer-CWA(Q,D) are non-empty

Naïve Evaluation

fails for non-positive queries

Lives	name	city
	Alice	$\bot_1$
	John	$\bot_2$
	Mary	London

$Q = \exists x \exists y\,(\mathrm{Lives}(x,y) \land y \neq \mathrm{London})$

Q(D) = {John, Mary}

but Answer-CWA(Q,D) and Answer-OWA(Q,D) are empty

Data Complexity of BIQA-CWA

Theorem: For a fixed query Q ∈ **CQ**, BIQA-CWA[Q](**CQ**) is in LOGSPACE

Proof: since we can rely on naïve evaluation

Recap

- Incomplete databases - some values are missing (marked nulls)
- Closed/open world completions of an incomplete database
- Certain answers - true in all completions
- Querying incomplete databases is a hard problem in general - undecidable under OWA, and coNP-complete in data complexity under CWA
- But for CQs is in LOGSPACE in data complexity for both CWA and OWA - this is due to naïve evaluation