Probabilistic Databases: The Dichotomy of PQE

Antoine Amarilli

TELECOM Paris

Research goal: Understanding the complexity of PQE

What is the complexity of $\mathrm{PQE}(Q)$ depending on the query Q?

Research goal: Understanding the complexity of PQE

What is the complexity of $\mathrm{PQE}(Q)$ depending on the query Q?

$\rightarrow$ Recall that we study **data complexity**, i.e., Q is **fixed** and the input is the **data**

Research goal: Understanding the complexity of PQE

What is the complexity of $\mathrm{PQE}(Q)$ depending on the query Q?

$\rightarrow$ Recall that we study **data complexity**, i.e., Q is **fixed** and the input is the **data**

For example:

- For $Q : R(x)$ the problem is **easy** (PTIME)
- For $Q : R(x), S(x, y), T(y)$ the problem is **hard** (#P-hard)

Research goal: Understanding the complexity of PQE

What is the complexity of $\mathrm{PQE}(Q)$ depending on the query Q?

$\rightarrow$ Recall that we study **data complexity**, i.e., Q is **fixed** and the input is the **data**

For example:

- For $Q : R(x)$ the problem is **easy** (PTIME)
- For $Q : R(x), S(x, y), T(y)$ the problem is **hard** (#P-hard)

We will present the **dichotomy** of [Dalvi and Suciu, 2007, Dalvi and Suciu, 2012]:

Research goal: Understanding the complexity of PQE

What is the complexity of $\mathrm{PQE}(Q)$ depending on the query Q?

$\rightarrow$ Recall that we study **data complexity**, i.e., Q is **fixed** and the input is the **data**

For example:

- For $Q : R(x)$ the problem is **easy** (PTIME)
- For $Q : R(x), S(x, y), T(y)$ the problem is **hard** (#P-hard)

We will present the **dichotomy** of [Dalvi and Suciu, 2007, Dalvi and Suciu, 2012]:

- Small dichotomy: **conjunctive queries** that are **self-join-free** and **arity-two**

Research goal: Understanding the complexity of PQE

What is the complexity of $\mathrm{PQE}(Q)$ depending on the query Q?

$\rightarrow$ Recall that we study **data complexity**, i.e., Q is **fixed** and the input is the **data**

For example:

- For $Q : R(x)$ the problem is **easy** (PTIME)
- For $Q : R(x), S(x, y), T(y)$ the problem is **hard** (#P-hard)

We will present the **dichotomy** of [Dalvi and Suciu, 2007, Dalvi and Suciu, 2012]:

- Small dichotomy: **conjunctive queries** that are **self-join-free** and **arity-two**
- Large dichotomy: arbitrary **unions of conjunctive queries**

Research goal: Understanding the complexity of PQE

What is the complexity of $\mathrm{PQE}(Q)$ depending on the query Q?

$\rightarrow$ Recall that we study **data complexity**, i.e., Q is **fixed** and the input is the **data**

For example:

- For $Q : R(x)$ the problem is **easy** (PTIME)
- For $Q : R(x), S(x, y), T(y)$ the problem is **hard** (#P-hard)

We will present the **dichotomy** of [Dalvi and Suciu, 2007, Dalvi and Suciu, 2012]:

- Small dichotomy: **conjunctive queries** that are **self-join-free** and **arity-two**
- Large dichotomy: arbitrary **unions of conjunctive queries**

Result of the form:

if Q *has a certain form* then $\mathrm{PQE}(Q)$ is in PTIME, otherwise it is #P-hard

The “small” Dalvi and Suciu dichotomy

- Conjunctive query (CQ): existentially quantified conjunction of atoms

The "small" Dalvi and Suciu dichotomy

- Conjunctive query (CQ): existentially quantified conjunction of atoms
- Arity-two: all relations are binary
  - We represent the queries as graphs: $R(x, y), S(y, z)$ is $x \longrightarrow y \longrightarrow z$

The "small" Dalvi and Suciu dichotomy

- Conjunctive query (CQ): existentially quantified conjunction of atoms
- Arity-two: all relations are binary
  - We represent the queries as graphs: $R(x, y), S(y, z)$ is $x \longrightarrow y \longrightarrow z$
- Self-join-free CQ: only one edge of each color (no repeated color)

The "small" Dalvi and Suciu dichotomy

- **Conjunctive query (CQ)**: existentially quantified conjunction of atoms
- **Arity-two**: all relations are binary
  - We represent the queries as **graphs**: $R(x, y), S(y, z)$ is $x \xrightarrow{\quad} y \xrightarrow{\quad} z$
- **Self-join-free CQ**: only one edge of each color (no repeated color)

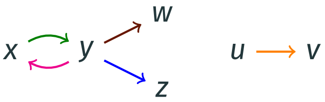
Theorem ([Dalvi and Suciu, 2007])

Let Q be an arity-two self-join-free CQ:

- *If Q is a **conjunction of stars**, then $\mathrm{PQE}(Q)$ is in **PTIME***
- *Otherwise, $\mathrm{PQE}(Q)$ is **#P-hard***

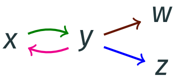
Conjunction of stars

- A **star** is a CQ with a **separator variable** that occurs in all edges
- A **conjunction of stars** is a conjunction of one or several stars



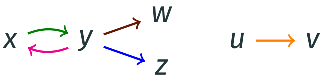
The following is **not a star**: $x \longrightarrow y \longrightarrow z \longrightarrow w$

Proving the small dichotomy (upper bound, 1)

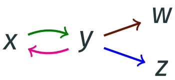


How to solve $\mathrm{PQE}(Q)$ for Q a conjunction of stars?

Proving the small dichotomy (upper bound, 1)



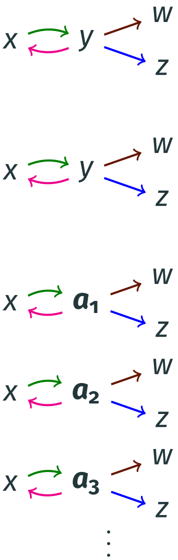
How to solve $\mathrm{PQE}(\boldsymbol{Q})$ for $\boldsymbol{Q}$ a conjunction of stars?



- We consider each connected component separately

$\rightarrow$ **Independent conjunction** over the connected components

Proving the small dichotomy (upper bound, 1)



$u \longrightarrow v$

How to solve $\mathrm{PQE}(Q)$ for Q a conjunction of stars?

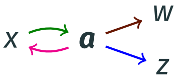
- We consider each connected component separately

$\rightarrow$ **Independent conjunction** over the connected components

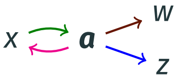
- We can test all possible values of the **separator variable**

$\rightarrow$ **Independent disjunction** over the values of the separator

Proving the small dichotomy (upper bound, 2)



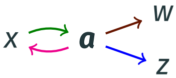
Proving the small dichotomy (upper bound, 2)



- For every match, we consider every **other variable** separately

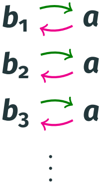
→ **Independent conjunction** over the variables

Proving the small dichotomy (upper bound, 2)



- For every match, we consider every **other variable** separately

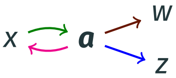
→ **Independent conjunction** over the variables



- We consider every value for the **other variable**

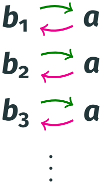
→ **Independent disjunction** over the possible assignments

Proving the small dichotomy (upper bound, 2)



- For every match, we consider every **other variable** separately

→ **Independent conjunction** over the variables



- We consider every value for the **other variable**

→ **Independent disjunction** over the possible assignments

- We consider every fact

→ **Independent conjunction** over the facts

→ Just **read the probability** of the ground fact $R(b, a)$.

Proving the small dichotomy (lower bound, 1)

Every arity-two self-join-free CQ which is not a conjunction of stars contains a pattern essentially like:

$$x \longrightarrow y \longrightarrow z \longrightarrow w$$

Proving the small dichotomy (lower bound, 1)

Every arity-two self-join-free CQ which is **not a conjunction of stars** contains a pattern essentially like:

$$x \longrightarrow y \longrightarrow z \longrightarrow w$$

We can **add facts with probability 1** to instances so the other facts are always satisfied, and focus on **only these three facts**

$\rightarrow$ **Let us show #P-hardness of this query**

Proving the small dichotomy (lower bound, 2)

Let us show that $\mathrm{PQE}(Q)$ is **#P-hard** for the CQ $Q: \; x \longrightarrow y \longrightarrow z \longrightarrow w$

Proving the small dichotomy (lower bound, 2)

Let us show that $\mathrm{PQE}(Q)$ is **#P-hard** for the CQ Q : $x \longrightarrow y \longrightarrow z \longrightarrow w$

- Reduce from the problem of **counting satisfying valuations** of a Boolean formula
  - e.g., given $(x \vee y) \wedge z$, compute that it has 3 satisfying valuations

Proving the small dichotomy (lower bound, 2)

Let us show that $\mathrm{PQE}(Q)$ is **#P-hard** for the CQ Q : $x \longrightarrow y \longrightarrow z \longrightarrow w$

- Reduce from the problem of **counting satisfying valuations** of a Boolean formula
  - e.g., given $(x \vee y) \wedge z$, compute that it has 3 satisfying valuations
- This problem is already **#P-hard** for so-called **PP2DNF formulas**:

Proving the small dichotomy (lower bound, 2)

Let us show that $\mathrm{PQE}(Q)$ is **#P-hard** for the CQ Q : $x \longrightarrow y \longrightarrow z \longrightarrow w$

- Reduce from the problem of **counting satisfying valuations** of a Boolean formula
  - e.g., given $(x \lor y) \land z$, compute that it has 3 satisfying valuations

- This problem is already **#P-hard** for so-called **PP2DNF formulas**:
  - **Positive** (no negation) and **Partitioned variables**: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$

Proving the small dichotomy (lower bound, 2)

Let us show that $\mathrm{PQE}(Q)$ is **#P-hard** for the CQ $Q: x \longrightarrow y \longrightarrow z \longrightarrow w$

- Reduce from the problem of **counting satisfying valuations** of a Boolean formula
  - e.g., given $(x \vee y) \wedge z$, compute that it has 3 satisfying valuations
- This problem is already **#P-hard** for so-called **PP2DNF formulas**:
  - **Positive** (no negation) and **Partitioned variables**: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$
  - **2-DNF**: disjunction of clauses like $X_i \wedge Y_j$

Proving the small dichotomy (lower bound, 2)

Let us show that $\mathrm{PQE}(Q)$ is **#P-hard** for the CQ Q : $x \xrightarrow{\quad} y \xrightarrow{\quad} z \xrightarrow{\quad} w$

- Reduce from the problem of **counting satisfying valuations** of a Boolean formula
  - e.g., given $(x \lor y) \land z$, compute that it has 3 satisfying valuations
- This problem is already **#P-hard** for so-called **PP2DNF formulas**:
  - **Positive** (no negation) and **Partitioned variables**: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$
  - **2-DNF**: disjunction of clauses like $X_i \land Y_j$
- Example: $\phi : (X_1 \land Y_1) \lor (X_1 \land Y_2) \lor (X_2 \land Y_2) \lor (X_3 \land Y_1) \lor (X_3 \land Y_2)$

Proving the small dichotomy (lower bound, 3)

Reduce from #PP2DNF to $\mathrm{PQE}(Q)$ for CQ Q : $x \longrightarrow y \longrightarrow z \longrightarrow w$

Example: $\phi : (X_1 \land Y_1) \lor (X_1 \land Y_2) \lor (X_2 \land Y_2) \lor (X_3 \land Y_1) \lor (X_3 \land Y_2)$

Proving the small dichotomy (lower bound, 3)

Reduce from #PP2DNF to $\mathrm{PQE}(Q)$ for CQ Q : $x \longrightarrow y \longrightarrow z \longrightarrow w$

Example: $\phi : (X_1 \wedge Y_1) \vee (X_1 \wedge Y_2) \vee (X_2 \wedge Y_2) \vee (X_3 \wedge Y_1) \vee (X_3 \wedge Y_2)$

Build an instance I_ϕ from ϕ:

Proving the small dichotomy (lower bound, 3)

Reduce from **#PP2DNF** to $\mathrm{PQE}(Q)$ for CQ Q : $x \longrightarrow y \longrightarrow z \longrightarrow w$

Example: $\phi : (X_1 \wedge Y_1) \vee (X_1 \wedge Y_2) \vee (X_2 \wedge Y_2) \vee (X_3 \wedge Y_1) \vee (X_3 \wedge Y_2)$

Build an **instance** I_ϕ from ϕ:

$$a'_1 \xrightarrow{1/2} a_1$$

$$a'_2 \xrightarrow{1/2} a_2$$

$$a'_3 \xrightarrow{1/2} a_3$$

Proving the small dichotomy (lower bound, 3)

Reduce from **#PP2DNF** to $\mathrm{PQE}(Q)$ for CQ Q : $x \longrightarrow y \longrightarrow z \longrightarrow w$

Example: $\phi : (X_1 \wedge Y_1) \vee (X_1 \wedge Y_2) \vee (X_2 \wedge Y_2) \vee (X_3 \wedge Y_1) \vee (X_3 \wedge Y_2)$

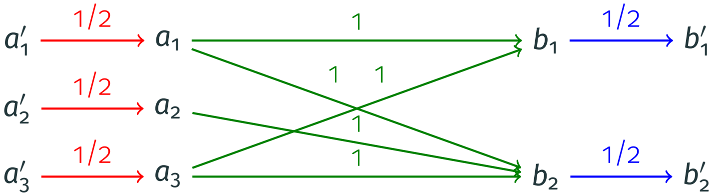
Build an **instance** I_ϕ from ϕ:

Proving the small dichotomy (lower bound, 3)

Reduce from **#PP2DNF** to $\mathrm{PQE}(Q)$ for CQ Q : $x \xrightarrow{} y \xrightarrow{} z \xrightarrow{} w$

Example: $\phi : (X_1 \wedge Y_1) \vee (X_1 \wedge Y_2) \vee (X_2 \wedge Y_2) \vee (X_3 \wedge Y_1) \vee (X_3 \wedge Y_2)$

Build an **instance** I_ϕ from ϕ:

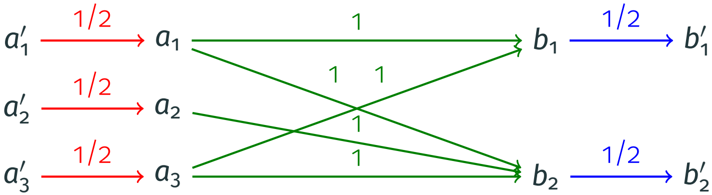


Proving the small dichotomy (lower bound, 3)

Reduce from #PP2DNF to $\mathrm{PQE}(Q)$ for CQ Q : $x \longrightarrow y \longrightarrow z \longrightarrow w$

Example: $\phi : (X_1 \wedge Y_1) \vee (X_1 \wedge Y_2) \vee (X_2 \wedge Y_2) \vee (X_3 \wedge Y_1) \vee (X_3 \wedge Y_2)$

Build an instance I_ϕ from ϕ:



Idea:

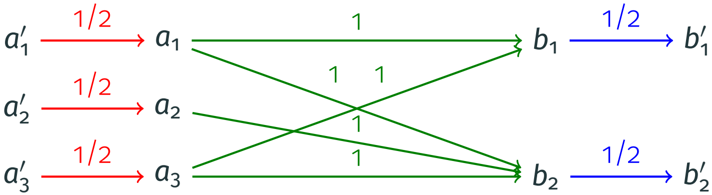
- **Valuations** of ϕ correspond to **possible worlds** of I_ϕ

Proving the small dichotomy (lower bound, 3)

Reduce from **#PP2DNF** to $\mathrm{PQE}(Q)$ for CQ Q : $x \longrightarrow y \longrightarrow z \longrightarrow w$

Example: $\phi : (X_1 \wedge Y_1) \vee (X_1 \wedge Y_2) \vee (X_2 \wedge Y_2) \vee (X_3 \wedge Y_1) \vee (X_3 \wedge Y_2)$

Build an **instance** I_ϕ from ϕ:



Idea:

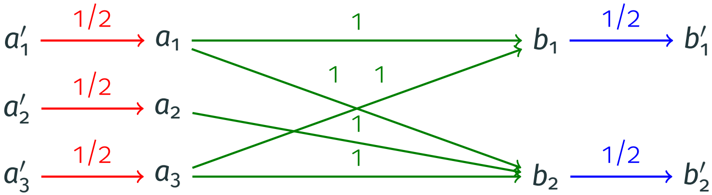
- **Valuations** of ϕ correspond to **possible worlds** of I_ϕ
- A valuation **satisfies** ϕ iff the corresponding possible world **satisfies** Q

Proving the small dichotomy (lower bound, 3)

Reduce from **#PP2DNF** to $\mathrm{PQE}(Q)$ for CQ Q : $x \longrightarrow y \longrightarrow z \longrightarrow w$

Example: $\phi : (X_1 \wedge Y_1) \vee (X_1 \wedge Y_2) \vee (X_2 \wedge Y_2) \vee (X_3 \wedge Y_1) \vee (X_3 \wedge Y_2)$

Build an **instance** I_ϕ from ϕ:



Idea:

- **Valuations** of ϕ correspond to **possible worlds** of I_ϕ
- A valuation **satisfies** ϕ iff the corresponding possible world **satisfies** Q

→ The **probability** of Q on I_ϕ is the **number of accepting valuations of** ϕ, divided by the number of valuations ($2^{-|\mathrm{Vars}|}$)

Extending beyond arity-two (1)

How can we extend beyond arity-two queries?

Theorem ([Dalvi and Suciu, 2007])

Let Q be a ~~arity-two~~ self-join-free CQ:

- *If Q is ~~a conjunction of stars~~ hierarchical, then $\mathrm{PQE}(Q)$ is in PTIME*
- *Otherwise, $\mathrm{PQE}(Q)$ is #P-hard*

Extending beyond arity-two (2)

Class of Hierarchical CQs defined inductively:

- A query with no variables is hierarchical

Extending beyond arity-two (2)

Class of Hierarchical CQs defined inductively:

- A query with no variables is hierarchical
- A conjunction of hierarchical connected components is hierarchical

Extending beyond arity-two (2)

Class of Hierarchical CQs defined inductively:

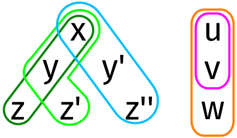
- A query with no variables is hierarchical
- A conjunction of hierarchical connected components is hierarchical
- Induction case: for a connected CQ:
  - It must have a separator variable occurring in all atoms
  - If we remove this separator variable, the query must be hierarchical

Extending beyond arity-two (2)

Class of Hierarchical CQs defined inductively:

- A query with **no variables** is hierarchical
- A **conjunction** of hierarchical connected components is hierarchical
- Induction case: for a connected CQ:
  - It must have a **separator variable** occurring in all atoms
  - If we remove this separator variable, the query must be hierarchical

$$\exists x\,(\exists y\,(\exists z\; R_1(x,y,z)) \wedge (\exists z'\; R_2(x,y,z'))) \wedge (\exists y' \exists z''\; R_3(x,y',z''))$$
$$\wedge\,(\exists u\,(\exists v\; R_4(u,v)) \wedge (\exists w\; R_5(u,v,w)))$$



Extending beyond arity-two (3)

How does the proof change?

Extending beyond arity-two (3)

How does the proof change?

- **Upper bound**: we can generalize the algorithm
  - **Independent AND** of connected components
  - **Independent OR** of possible choices for the separator variable
  - **Both cases use self-join-freeness!**

Extending beyond arity-two (3)

How does the proof change?

- **Upper bound**: we can generalize the algorithm
  - **Independent AND** of connected components
  - **Independent OR** of possible choices for the separator variable
  - **Both cases use self-join-freeness!**
- **Lower bound**: a non-hierarchical expression contains a pattern like

$$x \longrightarrow y \longrightarrow z \longrightarrow w$$

Extending beyond arity-two (3)

How does the proof change?

- **Upper bound**: we can generalize the algorithm
  - **Independent AND** of connected components
  - **Independent OR** of possible choices for the separator variable
  - **Both cases use self-join-freeness!**
- **Lower bound**: a non-hierarchical expression contains a pattern like

$$x \longrightarrow y \longrightarrow z \longrightarrow w$$

Via **equivalent characterization**: a non-hierarchical query has two variables x and y and:

- One atom containing **x and y**
- One atom containing **x but not y**
- One atom containing **y but not x**

The “big” Dalvi and Suciu dichotomy

Full dichotomy on the unions of conjunctive queries (UCQs):

Theorem ([Dalvi and Suciu, 2012])

Let Q *be a UCQ:*

- *If* Q *is handled by a complicated algorithm then* $\mathrm{PQE}(Q)$ *is in PTIME*
- *Otherwise,* $\mathrm{PQE}(Q)$ *is #P-hard*

The "big" Dalvi and Suciu dichotomy

Full dichotomy on the **unions of conjunctive queries** (UCQs):

Theorem ([Dalvi and Suciu, 2012])

Let Q be a UCQ:

- *If Q is handled by a complicated algorithm then $\mathrm{PQE}(Q)$ is in **PTIME***
- *Otherwise, $\mathrm{PQE}(Q)$ is **#P-hard***

This result is **far more challenging**:

- **Upper bound:**
  - an algorithm generalizing the previous case with **inclusion-exclusion**
  - many **unpleasant details** (e.g., a ranking transformation)

The "big" Dalvi and Suciu dichotomy

Full dichotomy on the **unions of conjunctive queries** (UCQs):

Theorem ([Dalvi and Suciu, 2012])

Let Q be a UCQ:

- *If Q is handled by a complicated algorithm then $\mathrm{PQE}(Q)$ is in **PTIME***
- *Otherwise, $\mathrm{PQE}(Q)$ is **#P-hard***

This result is **far more challenging**:

- **Upper bound:**
  - an algorithm generalizing the previous case with **inclusion-exclusion**
  - many **unpleasant details** (e.g., a ranking transformation)
- **Lower bound:** hardness proof on minimal cases where the algorithm does not work (very challenging)

References i

Dalvi, N. and Suciu, D. (2007).
The dichotomy of conjunctive queries on probabilistic structures.
In *Proc. PODS*.

Dalvi, N. and Suciu, D. (2012).
The dichotomy of probabilistic inference for unions of conjunctive queries.
J. ACM, 59(6).