# Probabilistic Databases: Models and PQE

Antoine Amarilli

# Relational model by example

Guest

| id | name | email |
|---|---|---|
| 1 | John Smith | john.smith@gmail.com |
| 2 | Alice Black | alice@black.name |
| 3 | John Smith | john.smith@ens.fr |

Reservation

| id | guest | room | arrival | nights |
|---|---|---|---|---|
| 1 | 1 | 504 | 2022-01-01 | 5 |
| 2 | 2 | 107 | 2022-01-10 | 3 |
| 3 | 3 | 302 | 2022-01-15 | 6 |
| 4 | 2 | 504 | 2022-01-15 | 2 |
| 5 | 2 | 107 | 2022-01-30 | 1 |

# Relations and databases

Formally:

- A **database schema** $\mathcal{D}$ maps each **relation name** to an **arity** (we add **attribute names** in our examples)

# Relations and databases

Formally:

- A **database schema** $\mathcal{D}$ maps each **relation name** to an **arity** (we add **attribute names** in our examples)
- A **database instance** over database schema $\mathcal{D}$ maps each **relation name** $R$ of $\mathcal{D}$ with arity $k$ to a set of $k$-tuples

# Relations and databases

Formally:

- A **database schema** $\mathcal{D}$ maps each **relation name** to an **arity** (we add **attribute names** in our examples)
- A **database instance** over database schema $\mathcal{D}$ maps each **relation name** $R$ of $\mathcal{D}$ with arity $k$ to a set of $k$-tuples

We can write tuples as **table rows** or as **ground facts**:

**Guest**

| id | name | email |
|---|---|---|
| 1 | John Smith | john.smith@gmail.com |
| 2 | Alice Black | alice@black.name |
| 3 | John Smith | john.smith@ens.fr |

Guest(1, John Smith, john.smith@gmail.com),
Guest(2, Alice Black, alice@black.name),
Guest(3, John Smith, john.smith@ens.fr)

# Queries

- A **query** is an arbitrary **function** over database instances over a fixed schema $\mathcal{D}$
- We only study **Boolean queries**, i.e., queries returning only **true** or **false**

# Queries

- A **query** is an arbitrary **function** over database instances over a fixed schema $\mathcal{D}$
- We only study **Boolean queries**, i.e., queries returning only **true** or **false**
- Example of query languages:
  - **Conjunctive queries** (CQ)
    - $\exists \bigwedge \cdots$: existentially quantified conjunctions of atoms
    - $Q : \exists x\, y\, z\, x'\, y'\ Guest(x, y, z) \land Guest(x', y', z)$

# Queries

- A **query** is an arbitrary **function** over database instances over a fixed schema $\mathcal{D}$
- We only study **Boolean queries**, i.e., queries returning only **true** or **false**
- Example of query languages:
  - Conjunctive queries (CQ)
    - $\exists \bigwedge \cdots$: existentially quantified conjunctions of atoms
    - $Q : \exists x\, y\, z\, x'\, y'\ Guest(x, y, z) \wedge Guest(x', y', z)$
  - Unions of conjunctive queries (UCQ)
    - $\bigcup \exists \bigwedge \cdots$: unions of CQs

# Queries

- A **query** is an arbitrary **function** over database instances over a fixed schema $\mathcal{D}$
- We only study **Boolean queries**, i.e., queries returning only **true** or **false**
- Example of query languages:
  - **Conjunctive queries** (CQ)
    - $\exists \bigwedge \cdots$: existentially quantified conjunctions of atoms
    - $Q : \exists x\, y\, z\, x'\, y'\ Guest(x, y, z) \wedge Guest(x', y', z)$
  - **Unions of conjunctive queries** (UCQ)
    - $\bigcup \exists \bigwedge \cdots$: unions of **CQs**
  - First-Order logic (FO)
  - **Monadic-Second Order** logic (MSO)

# TID

# Tuple-independent databases (TID)

- The **simplest** model: tuple-independent databases
- Annotate each **instance fact** with a **probability**

# Tuple-independent databases (TID)

- The **simplest** model: tuple-independent databases
- Annotate each **instance fact** with a **probability**

| date | teacher |
|---|---|
| 08 | Diego |
| 09 | Paolo |
| 09 | Floris |

# Tuple-independent databases (TID)

- The **simplest** model: tuple-independent databases
- Annotate each **instance fact** with a **probability**

| date | teacher | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

# Tuple-independent databases (TID)

- The **simplest** model: tuple-independent databases
- Annotate each **instance fact** with a **probability**

| **date** | **teacher** | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

→ Assume **independence** between facts

# Semantics of TID

- Each fact is kept or discarded with the indicated probability
- Probabilistic choices are independent across facts

# Semantics of TID

- Each fact is kept or discarded with the indicated probability
- Probabilistic choices are independent across facts

| **date** | **teacher** | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

# Semantics of TID

- Each fact is kept or discarded with the indicated probability
- Probabilistic choices are independent across facts

| date | teacher | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

| date | teacher |
|---|---|

# Semantics of TID

- Each fact is **kept** or **discarded** with the indicated probability
- Probabilistic choices are **independent** across facts

| date | teacher | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

| date | teacher |
|---|---|
| 08 | Diego |

# Semantics of TID

- Each fact is **kept** or **discarded** with the indicated probability
- Probabilistic choices are **independent** across facts

| date | teacher | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

| date | teacher |
|---|---|
| 08 | Diego |
| 09 | Paolo |

# Semantics of TID

- Each fact is **kept** or **discarded** with the indicated probability
- Probabilistic choices are **independent** across facts

| date | teacher | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

| date | teacher |
|---|---|
| 08 | Diego |
| 09 | Paolo |
| 09 | Floris |

# Semantics of TID

- Each fact is **kept** or **discarded** with the indicated probability
- Probabilistic choices are **independent** across facts

| **date** | **teacher** | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

| **date** | **teacher** |
|---|---|
| 08 | Diego |
| 09 | Paolo |
| 09 | Floris |

What's the **probability** of this possible world?

## Semantics of TID

- Each fact is **kept** or **discarded** with the indicated probability
- Probabilistic choices are **independent** across facts

| date | teacher | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

| date | teacher |
|---|---|
| 08 | Diego |
| 09 | Paolo |
| 09 | Floris |

What's the **probability** of this possible world?

90%

## Semantics of TID

- Each fact is kept or discarded with the indicated probability
- Probabilistic choices are independent across facts

| date | teacher | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

| date | teacher |
|---|---|
| 08 | Diego |
| 09 | Paolo |
| 09 | Floris |

What's the probability of this possible world?

$$90\% \times$$

# Semantics of TID

- Each fact is **kept** or **discarded** with the indicated probability
- Probabilistic choices are **independent** across facts

| date | teacher | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

| date | teacher |
|---|---|
| 08 | Diego |
| 09 | Paolo |
| 09 | Floris |

What's the **probability** of this possible world?

$$90\% \times (100\% - 80\%)$$

# Semantics of TID

- Each fact is **kept** or **discarded** with the indicated probability
- Probabilistic choices are **independent** across facts

| date | teacher | |
|---|---|---|
| 08 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

| date | teacher |
|---|---|
| 08 | Diego |
| 09 | Paolo |
| 09 | Floris |

What's the **probability** of this possible world?

$$90\% \times (100\% - 80\%) \times 70\%$$

# Getting a probability distribution

The semantics of a TID $I$ is a probability distribution on (non-probabilistic) databases...

$\rightarrow$ the possible worlds are the subsets of facts of $I$

# Getting a probability distribution

The semantics of a TID $I$ is a probability distribution on (non-probabilistic) databases...

$\rightarrow$ the possible worlds are the subsets of facts of $I$

  $\rightarrow$ always keeping facts with probability 1

# Getting a probability distribution

The semantics of a TID $I$ is a probability distribution on (non-probabilistic) databases...

$\rightarrow$ the possible worlds are the subsets of facts of $I$

$\rightarrow$ always keeping facts with probability 1

Formally, for a TID $I$, the probability of $J \subseteq I$ is:

# Getting a probability distribution

The semantics of a TID $I$ is a probability distribution on (non-probabilistic) databases...

$\rightarrow$ the possible worlds are the subsets of facts of $I$

$\rightarrow$ always keeping facts with probability 1

Formally, for a TID $I$, the probability of $J \subseteq I$ is:

- product of $\Pr(F)$ for each fact $F$ kept in $J$
- product of $1 - \Pr(F)$ for each fact $F$ not kept in $J$

# Is it a probability distribution?

Do the probabilities of the possible words always sum to 1?

# Is it a probability distribution?

Do the probabilities of the possible words always sum to 1?

- Let $N$ be the number of facts
- There are $2^N$ possible worlds

# Is it a probability distribution?

Do the probabilities of the possible words always **sum to 1**?

- Let $N$ be the **number of facts**
- There are $\mathbf{2}^N$ **possible worlds**
- The probability of a possible world is a product which involves a factor $\Pr(F_i)$ or $1 - \Pr(F_i)$ for each fact $F_1, \ldots, F_N$

# Is it a probability distribution?

Do the probabilities of the possible words always **sum to 1**?

- Let $N$ be the **number of facts**
- There are $2^N$ **possible worlds**
- The probability of a possible world is a product which involves a factor $\Pr(F_i)$ or $1 - \Pr(F_i)$ for each fact $F_1, \ldots, F_N$

$\rightarrow$ The sum of these probabilities is the result of **expanding** the expression:
$(\Pr(F_1) + (1 - \Pr(F_1))) \times \cdots \times (\Pr(F_N) + (1 - \Pr(F_N)))$

# Is it a probability distribution?

Do the probabilities of the possible words always **sum to 1**?

- Let $N$ be the **number of facts**
- There are $2^N$ **possible worlds**
- The probability of a possible world is a product which involves a factor $\Pr(F_i)$ or $1 - \Pr(F_i)$ for each fact $F_1, \ldots, F_N$

$\rightarrow$ The sum of these probabilities is the result of **expanding** the expression: $(\Pr(F_1) + (1 - \Pr(F_1))) \times \cdots \times (\Pr(F_N) + (1 - \Pr(F_N)))$

- All factors are **equal to 1**, so the probabilities **sum to 1**

# Expressiveness of TID

Can we represent all probabilistic instances with TID?

# Expressiveness of TID

Can we represent **all** probabilistic instances with TID?

*"The class is taught by Jane or Joe or no one but **not both**"*

# Expressiveness of TID

Can we represent **all** probabilistic instances with TID?

*"The class is taught by Jane or Joe or no one but **not both**"*

$U_1$

| **teacher** |
| --- |
| Jane |

$\pi(U_1) = 80\%$

# Expressiveness of TID

Can we represent **all** probabilistic instances with TID?

*"The class is taught by Jane or Joe or no one but **not both**"*

$U_1$

| **teacher** |
| --- |
| Jane |
| $\pi(U_1) = 80\%$ |

$U_2$

| **teacher** |
| --- |
| Joe |
| $\pi(U_2) = 10\%$ |

# Expressiveness of TID

Can we represent **all** probabilistic instances with TID?

*"The class is taught by Jane or Joe or no one but **not both**"*

$U_1$

| **teacher** |
|---|
| Jane |
| $\pi(U_1) = 80\%$ |

$U_2$

| **teacher** |
|---|
| Joe |
| $\pi(U_2) = 10\%$ |

$U_3$

| **teacher** |
|---|
| |
| $\pi(U_3) = 10\%$ |

# Expressiveness of TID

Can we represent **all** probabilistic instances with TID?

*"The class is taught by Jane or Joe or no one but **not both**"*

$U_1$

| teacher |
|---|
| Jane |

$\pi(U_1) = 80\%$

$U_2$

| teacher |
|---|
| Joe |

$\pi(U_2) = 10\%$

$U_3$

| teacher |
|---|
| |

$\pi(U_3) = 10\%$

| teacher |
|---|
| Jane |
| Joe |

## Expressiveness of TID

Can we represent **all** probabilistic instances with TID?

*"The class is taught by Jane or Joe or no one but **not both**"*

$U_1$

| **teacher** |
|---|
| Jane |

$\pi(U_1) = 80\%$

$U_2$

| **teacher** |
|---|
| Joe |

$\pi(U_2) = 10\%$

$U_3$

| **teacher** |
|---|
| |

$\pi(U_3) = 10\%$

| **teacher** | |
|---|---|
| Jane | 10% |
| Joe | |

# Expressiveness of TID

Can we represent **all** probabilistic instances with TID?

*"The class is taught by Jane or Joe or no one but **not both**"*

$U_1$

| **teacher** |
|---|
| Jane |

$\pi(U_1) = 80\%$

$U_2$

| **teacher** |
|---|
| Joe |

$\pi(U_2) = 10\%$

$U_3$

| **teacher** |
|---|
| |

$\pi(U_3) = 10\%$

| **teacher** | |
|---|---|
| Jane | 10% |
| Joe | 80% |

# Expressiveness of TID

Can we represent **all** probabilistic instances with TID?

*"The class is taught by Jane or Joe or no one but **not both**"*

$U_1$

| **teacher** |
|---|
| Jane |

$\pi(U_1) = 80\%$

$U_2$

| **teacher** |
|---|
| Joe |

$\pi(U_2) = 10\%$

$U_3$

| **teacher** |
|---|
| |

$\pi(U_3) = 10\%$

| **teacher** | |
|---|---|
| Jane | 10% |
| Joe | 80% |

$\rightarrow$ We **cannot** forbid that both teach the class!

# BID

# Block-independent disjoint instances

- A more expressive framework than TID
- Call some attributes the key (underlined)

# Block-independent disjoint instances

- A more expressive framework than TID
- Call some attributes the key (underlined)

*U*

| day | time | teacher |
|---|---|---|
| 09 | AM | Paolo |
| 09 | AM | Floris |
| 09 | PM | Floris |
| 09 | PM | Paolo |

# Block-independent disjoint instances

- A more expressive framework than TID
- Call some attributes the key (underlined)

*U*

| day | time | teacher |
|---|---|---|
| 09 | AM | Paolo |
| 09 | AM | Floris |
| 09 | PM | Floris |
| 09 | PM | Paolo |

- The blocks are the sets of tuples with the same key

# Block-independent disjoint instances

- A **more expressive framework** than TID
- Call some attributes the **key** (underlined)

*U*

| day | time | teacher |
|---|---|---|
| 09 | AM | Paolo |
| 09 | AM | Floris |
| 09 | PM | Floris |
| 09 | PM | Paolo |

- The **blocks** are the sets of tuples with the same key
- Each **tuple** has a probability

# Block-independent disjoint instances

- A **more expressive framework** than TID
- Call some attributes the **key** (underlined)

*U*

| day | time | teacher | |
|---|---|---|---|
| 09 | AM | Paolo | 80% |
| 09 | AM | Floris | 10% |
| 09 | PM | Floris | 70% |
| 09 | PM | Paolo | 1% |

- The **blocks** are the sets of tuples with the same key
- Each **tuple** has a probability

# Block-independent disjoint instances

- A more expressive framework than TID
- Call some attributes the key (underlined)

*U*

| day | time | teacher | |
|---|---|---|---|
| 09 | AM | Paolo | 80% |
| 09 | AM | Floris | 10% |
| 09 | PM | Floris | 70% |
| 09 | PM | Paolo | 1% |

- The blocks are the sets of tuples with the same key
- Each tuple has a probability
- Probabilities must sum up to $\leq 1$ in each block

# BID semantics

*U*

| <u>day</u> | <u>time</u> | teacher | |
|---|---|---|---|
| 09 | AM | Paolo | 80% |
| 09 | AM | Floris | 10% |
| 09 | PM | Floris | 70% |
| 09 | PM | Paolo | 1% |

# BID semantics

*U*

| <u>day</u> | <u>time</u> | teacher | |
|---|---|---|---|
| 09 | AM | Paolo | 80% |
| 09 | AM | Floris | 10% |
| 09 | PM | Floris | 70% |
| 09 | PM | Paolo | 1% |

- For each **block**:

# BID semantics

*U*

| day | time | teacher | |
|---|---|---|---|
| 09 | AM | Paolo | 80% |
| 09 | AM | Floris | 10% |
| 09 | PM | Floris | 70% |
| 09 | PM | Paolo | 1% |

- For each **block**:
  - Pick **one** fact according to probabilities

# BID semantics

*U*

| <u>day</u> | <u>time</u> | teacher | |
|---|---|---|---|
| 09 | AM | Paolo | 80% |
| 09 | AM | Floris | 10% |
| 09 | PM | Floris | 70% |
| 09 | PM | Paolo | 1% |

- For each **block**:
  - Pick **one** fact according to probabilities
  - Possibly **no** fact if probabilities sum up to $< 1$

# BID semantics

| | | $U$ | |
|---|---|---|---|
| <u>day</u> | <u>time</u> | teacher | |
| 09 | AM | Paolo | 80% |
| 09 | AM | Floris | 10% |
| 09 | PM | Floris | 70% |
| 09 | PM | Paolo | 1% |

- For each **block**:
  - Pick **one** fact according to probabilities
  - Possibly **no** fact if probabilities sum up to $< 1$

$\rightarrow$ Do choices **independently** in each block

# BID semantics

| | | $U$ | | | | $U$ |
|---|---|---|---|---|---|---|
| day | time | teacher | | day | time | teacher |
| 09 | AM | Paolo | 80% | | | |
| 09 | AM | Floris | 10% | | | |
| 09 | PM | Floris | 70% | | | |
| 09 | PM | Paolo | 1% | | | |

- For each **block**:
  - Pick **one** fact according to probabilities
  - Possibly **no** fact if probabilities sum up to $< 1$

$\rightarrow$ Do choices **independently** in each block

# BID semantics

*U*

| day | time | teacher | |
|---|---|---|---|
| 09 | AM | Paolo | 80% |
| 09 | AM | Floris | 10% |
| 09 | PM | Floris | 70% |
| 09 | PM | Paolo | 1% |

*U*

| day | time | teacher |
|---|---|---|
| 09 | AM | Paolo |
| 09 | AM | Floris |

- For each **block**:
  - Pick **one** fact according to probabilities
  - Possibly **no** fact if probabilities sum up to $< 1$

$\rightarrow$ Do choices **independently** in each block

# BID semantics

| *U* | | | |
|---|---|---|---|
| <u>day</u> | <u>time</u> | teacher | |
| 09 | AM | Paolo | 80% |
| 09 | AM | Floris | 10% |
| 09 | PM | Floris | 70% |
| 09 | PM | Paolo | 1% |

| *U* | | |
|---|---|---|
| <u>day</u> | <u>time</u> | teacher |
| 09 | AM | Paolo |
| 09 | AM | Floris |
| 09 | PM | Floris |
| 09 | PM | Paolo |

- For each **block**:
  - Pick **one** fact according to probabilities
  - Possibly **no** fact if probabilities sum up to $< 1$

$\rightarrow$ Do choices **independently** in each block

# BID captures TID

- Each TID can be expressed as a BID...

# BID captures TID

- Each TID can be expressed as a BID...
  - → Take all attributes as key
  - → Each block contains a single fact

# BID captures TID

- Each TID can be expressed as a BID...
  - → Take all attributes as key
  - → Each block contains a single fact

*U*

| date | teacher | |
|---|---|---|
| 09 | Diego | 90% |
| 09 | Paolo | 80% |
| 09 | Floris | 70% |

# Expressiveness of BID

Can we represent all probabilistic instances with BID?

# Expressiveness of BID

Can we represent all probabilistic instances with BID?

*"The class is taught by exactly two among Diego, Paolo, Floris."*

## Expressiveness of BID

Can we represent **all** probabilistic instances with BID?

*"The class is taught by exactly two among Diego, Paolo, Floris."*

| $U_1$ |
| --- |
| **teacher** |
| Diego |
| Paolo |
| $\pi(U_1) = 80\%$ |

## Expressiveness of BID

Can we represent **all** probabilistic instances with BID?

*"The class is taught by exactly two among Diego, Paolo, Floris."*

$U_1$

| teacher |
|---|
| Diego |
| Paolo |
| $\pi(U_1) = 80\%$ |

$U_2$

| teacher |
|---|
| Diego |
| Floris |
| $\pi(U_2) = 10\%$ |

# Expressiveness of BID

Can we represent **all** probabilistic instances with BID?

*"The class is taught by exactly two among Diego, Paolo, Floris."*

| $U_1$ |
|---|
| **teacher** |
| Diego |
| Paolo |
| $\pi(U_1) = 80\%$ |

| $U_2$ |
|---|
| **teacher** |
| Diego |
| Floris |
| $\pi(U_2) = 10\%$ |

| $U_3$ |
|---|
| **teacher** |
| Paolo |
| Floris |
| $\pi(U_3) = 10\%$ |

# Expressiveness of BID

Can we represent **all** probabilistic instances with BID?

*"The class is taught by exactly two among Diego, Paolo, Floris."*

$U_1$

| **teacher** |
|---|
| Diego |
| Paolo |
| $\pi(U_1) = 80\%$ |

$U_2$

| **teacher** |
|---|
| Diego |
| Floris |
| $\pi(U_2) = 10\%$ |

$U_3$

| **teacher** |
|---|
| Paolo |
| Floris |
| $\pi(U_3) = 10\%$ |

$\rightarrow$ If **teacher** is a key **teacher**, then TID

# Expressiveness of BID

Can we represent **all** probabilistic instances with BID?

*"The class is taught by exactly two among Diego, Paolo, Floris."*

$U_1$

| **teacher** |
|---|
| Diego |
| Paolo |
| $\pi(U_1) = 80\%$ |

$U_2$

| **teacher** |
|---|
| Diego |
| Floris |
| $\pi(U_2) = 10\%$ |

$U_3$

| **teacher** |
|---|
| Paolo |
| Floris |
| $\pi(U_3) = 10\%$ |

→ If **teacher** is a key **teacher**, then TID

→ If **teacher** is not a key, then **only one fact**

# Expressiveness of BID

Can we represent **all** probabilistic instances with BID?

*"The class is taught by exactly two among Diego, Paolo, Floris."*

$U_1$

| **teacher** |
|---|
| Diego |
| Paolo |
| $\pi(U_1) = 80\%$ |

$U_2$

| **teacher** |
|---|
| Diego |
| Floris |
| $\pi(U_2) = 10\%$ |

$U_3$

| **teacher** |
|---|
| Paolo |
| Floris |
| $\pi(U_3) = 10\%$ |

$\rightarrow$ If **teacher** is a key **teacher**, then **TID**

$\rightarrow$ If **teacher** is not a key, then **only one fact**

$\rightarrow$ We **cannot represent** this probabilistic instance as a BID

# pc-tables

# Boolean c-tables

- Set of **Boolean variables** $x_1, x_2, \ldots$
- Each **fact** has a **condition**: Variables, Boolean operators

# Boolean c-tables

- Set of **Boolean variables** $x_1, x_2, \ldots$
- Each **fact** has a **condition**: Variables, Boolean operators

| date | teacher | room | |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \land \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \land x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \land \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \land x_1$ |

$x_1$ Jane is sick

$x_2$ Amphi B is available

# pc-tables

A (Boolean) pc-table is:

- a database $I$ where each tuple is annotated by a Boolean function on variables $x_i$
- a probability $p_i$ that each variable $x_i$ is true (assuming independence)

# pc-tables

A (Boolean) **pc-table** is:

- a database $I$ where each tuple is annotated by a **Boolean function** on variables $x_i$
- a **probability** $p_i$ that each variable $x_i$ is true (assuming independence)

**Formally:**

- A Boolean **valuation** $\nu$ of the variables maps each variable $x_i$ to $0$ or $1$

# pc-tables

A (Boolean) **pc-table** is:

- a database $I$ where each tuple is annotated by a **Boolean function** on variables $x_i$
- a **probability** $p_i$ that each variable $x_i$ is true (assuming independence)

**Formally:**

- A Boolean **valuation** $\nu$ of the variables maps each variable $x_i$ to 0 or 1
- The valuation $\nu$ defines a **possible world** $I_\nu$ of $I$ containing the tuples whose Boolean function evaluates to **true** under $\nu$

# pc-tables

A (Boolean) **pc-table** is:

- a database $I$ where each tuple is annotated by a **Boolean function** on variables $x_i$
- a **probability** $p_i$ that each variable $x_i$ is true (assuming independence)

**Formally:**

- A Boolean **valuation** $\nu$ of the variables maps each variable $x_i$ to 0 or 1
- The valuation $\nu$ defines a **possible world** $I_\nu$ of $I$ containing the tuples whose Boolean function evaluates to **true** under $\nu$
- The **probability** of a valuation $\nu$ is:

# pc-tables

A (Boolean) **pc-table** is:

- a database $I$ where each tuple is annotated by a **Boolean function** on variables $x_i$
- a **probability** $p_i$ that each variable $x_i$ is true (assuming independence)

**Formally:**

- A Boolean **valuation** $\nu$ of the variables maps each variable $x_i$ to 0 or 1
- The valuation $\nu$ defines a **possible world** $I_\nu$ of $I$ containing the tuples whose Boolean function evaluates to **true** under $\nu$
- The **probability** of a valuation $\nu$ is:
    - Product of the $p_i$ for the $x_i$ with $\nu(x_i) = 1$

# pc-tables

A (Boolean) **pc-table** is:

- a database $I$ where each tuple is annotated by a **Boolean function** on variables $x_i$
- a **probability** $p_i$ that each variable $x_i$ is true (assuming independence)

**Formally:**

- A Boolean **valuation** $\nu$ of the variables maps each variable $x_i$ to $0$ or $1$
- The valuation $\nu$ defines a **possible world** $I_\nu$ of $I$ containing the tuples whose Boolean function evaluates to **true** under $\nu$
- The **probability** of a valuation $\nu$ is:
  - Product of the $p_i$ for the $x_i$ with $\nu(x_i) = 1$
  - Product of the $1 - p_i$ for the $x_i$ with $\nu(x_i) = 0$

## pc-tables

A (Boolean) **pc-table** is:

- a database $I$ where each tuple is annotated by a **Boolean function** on variables $x_i$
- a **probability** $p_i$ that each variable $x_i$ is true (assuming independence)

**Formally:**

- A Boolean **valuation** $\nu$ of the variables maps each variable $x_i$ to 0 or 1
- The valuation $\nu$ defines a **possible world** $I_\nu$ of $I$ containing the tuples whose Boolean function evaluates to **true** under $\nu$
- The **probability** of a valuation $\nu$ is:
  - Product of the $p_i$ for the $x_i$ with $\nu(x_i) = 1$
  - Product of the $1 - p_i$ for the $x_i$ with $\nu(x_i) = 0$

$\rightarrow$ This is like TIDs

# pc-tables

A (Boolean) **pc-table** is:

- a database $I$ where each tuple is annotated by a **Boolean function** on variables $x_i$
- a **probability** $p_i$ that each variable $x_i$ is true (assuming independence)

**Formally:**

- A Boolean **valuation** $\nu$ of the variables maps each variable $x_i$ to 0 or 1
- The valuation $\nu$ defines a **possible world** $I_\nu$ of $I$ containing the tuples whose Boolean function evaluates to **true** under $\nu$
- The **probability** of a valuation $\nu$ is:
    - Product of the $p_i$ for the $x_i$ with $\nu(x_i) = 1$
    - Product of the $1 - p_i$ for the $x_i$ with $\nu(x_i) = 0$

  $\rightarrow$ This is like TIDs
- The **probability** of a possible world $J \subseteq I$ is the total probability of the valuations $\nu$ such that $I_\nu = J$

# pc-table example

| date | teacher | room | |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \land \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \land x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \land \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \land x_1$ |

# pc-table example

| date | teacher | room | |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \wedge \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \wedge x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \wedge \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \wedge x_1$ |

$x_1$ Jane is sick

$x_2$ Amphi B is available

# pc-table example

| date | teacher | room | |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \land \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \land x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \land \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \land x_1$ |

$x_1$ Jane is sick

$\rightarrow$ **Probability 10%**

$x_2$ Amphi B is available

$\rightarrow$ **Probability 20%**

# pc-table semantics example

| date | teacher | room | $x_1 : 10\%, x_2 : 20\%$ |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \land \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \land x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \land \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \land x_1$ |

# pc-table semantics example

| date | teacher | room | $x_1 : 10\%, x_2 : 20\%$ |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \land \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \land x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \land \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \land x_1$ |

- Take $\nu$ mapping $x_1$ to $0$ and $x_2$ to $1$

# pc-table semantics example

| date | teacher | room | $x_1 : 10\%, x_2 : 20\%$ |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \wedge \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \wedge x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \wedge \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \wedge x_1$ |

- Take $\nu$ mapping $x_1$ to $0$ and $x_2$ to $1$
- Probability of $\nu$:

# pc-table semantics example

| date | teacher | room | $x_1 : 10\%, x_2 : 20\%$ |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \wedge \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \wedge x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \wedge \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \wedge x_1$ |

- Take $\nu$ mapping $x_1$ to $0$ and $x_2$ to $1$
- Probability of $\nu$: $(100\% - 10\%) \times 20\% = 18\%$

# pc-table semantics example

| date | teacher | room | $x_1 : 10\%, x_2 : 20\%$ |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \land \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \land x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \land \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \land x_1$ |

- Take $\nu$ mapping $x_1$ to $0$ and $x_2$ to $1$
- Probability of $\nu$: $(100\% - 10\%) \times 20\% = 18\%$
- Evaluate the conditions

# pc-table semantics example

| date | teacher | room | $x_1 : 10\%, x_2 : 20\%$ |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \land \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \land x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \land \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \land x_1$ |

| date | teacher | room |
|---|---|---|
| 04 | Jane | Amphi A |
| 04 | Joe | Amphi A |
| 11 | Jane | Amphi B |
| 11 | Joe | Amphi B |
| 11 | Jane | Amphi C |
| 11 | Joe | Amphi C |

- Take $\nu$ mapping $x_1$ to **0** and $x_2$ to **1**
- Probability of $\nu$: $(100\% - 10\%) \times 20\% = 18\%$
- Evaluate the conditions

# pc-table semantics example

| date | teacher | room | $x_1 : 10\%, x_2 : 20\%$ |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \land \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \land x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \land \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \land x_1$ |

| date | teacher | room |
|---|---|---|
| 04 | Jane | Amphi A |
| 04 | Joe | Amphi A |
| 11 | Jane | Amphi B |
| 11 | Joe | Amphi B |
| 11 | Jane | Amphi C |
| 11 | Joe | Amphi C |

- Take $\nu$ mapping $x_1$ to 0 and $x_2$ to 1
- Probability of $\nu$: $(100\% - 10\%) \times 20\% = 18\%$
- Evaluate the conditions

$\rightarrow$ Probability of possible world: sum over the valuations

# pc-table semantics example

| date | teacher | room | $x_1 : 10\%, x_2 : 20\%$ |
|---|---|---|---|
| 04 | Jane | Amphi A | $\neg x_1$ |
| 04 | Joe | Amphi A | $x_1$ |
| 11 | Jane | Amphi B | $x_2 \wedge \neg x_1$ |
| 11 | Joe | Amphi B | $x_2 \wedge x_1$ |
| 11 | Jane | Amphi C | $\neg x_2 \wedge \neg x_1$ |
| 11 | Joe | Amphi C | $\neg x_2 \wedge x_1$ |

| date | teacher | room |
|---|---|---|
| 04 | Jane | Amphi A |
| 04 | Joe | Amphi A |
| 11 | Jane | Amphi B |
| 11 | Joe | Amphi B |
| 11 | Jane | Amphi C |
| 11 | Joe | Amphi C |

- Take $\nu$ mapping $x_1$ to $0$ and $x_2$ to $1$
- Probability of $\nu$: $(100\% - 10\%) \times 20\% = 18\%$
- Evaluate the conditions

$\rightarrow$ Probability of possible world: sum over the valuations

$\rightarrow$ Here: only this valuation, 18%

# Expressiveness of pc-tables

- pc-tables capture **TIDs**:
  - $\rightarrow$ Simply give each fact its own **probability value**

# Expressiveness of pc-tables

- pc-tables capture **TIDs**:
  - → Simply give each fact its own **probability value**
- pc-tables capture **BIDs**:
  - → Make a **decision tree** for every block

# Expressiveness of pc-tables

- pc-tables capture **TIDs**:
  - $\rightarrow$ Simply give each fact its own **probability value**
- pc-tables capture **BIDs**:
  - $\rightarrow$ Make a **decision tree** for every block
- In fact pc-tables can express **arbitrary probability distributions**

# Expressiveness of pc-tables

- pc-tables capture **TIDs**:
  - $\rightarrow$ Simply give each fact its own **probability value**
- pc-tables capture **BIDs**:
  - $\rightarrow$ Make a **decision tree** for every block
- In fact pc-tables can express **arbitrary probability distributions**
- Further, they are a **strong representation system**: the union, product, etc., of two pc-tables, can be easily represented as a pc-table

# Expressiveness of pc-tables

- pc-tables capture **TIDs**:
  $\rightarrow$ Simply give each fact its own **probability value**

- pc-tables capture **BIDs**:
  $\rightarrow$ Make a **decision tree** for every block

- In fact pc-tables can express **arbitrary probability distributions**

- Further, they are a **strong representation system**: the union, product, etc., of two pc-tables, can be easily represented as a pc-table

Yet, in the rest of the class, we focus on **TIDs** $\rightarrow$ easier to characterize tractable queries

# PQE

# Query evaluation on probabilistic databases (PQE)

How can we evaluate a query $Q$ over a probabilistic database $D$?

# Query evaluation on probabilistic databases (PQE)

How can we evaluate a query $Q$ over a probabilistic database $D$?

- Probability that $Q$ holds over $D$:

$$\Pr(D \models Q) = \sum_{\substack{D' \subseteq D \\ D' \models Q}} \Pr(D')$$

- **Intuitively:** the probability that $Q$ holds is the probability of drawing a possible world $D' \subseteq D$ which satisfies $Q$

# Query evaluation on probabilistic databases (PQE)

How can we evaluate a query $Q$ over a probabilistic database $D$?

- Probability that $Q$ holds over $D$:

$$\Pr(D \models Q) = \sum_{\substack{D' \subseteq D \\ D' \models Q}} \Pr(D')$$

- **Intuitively:** the probability that $Q$ holds is the probability of drawing a possible world $D' \subseteq D$ which satisfies $Q$

**Probabilistic query evaluation (PQE)** problem for a query $Q$ over TIDs: given a TID, compute the probability that $Q$ holds

# Example of PQE on TID

| name | position | city | classification | prob |
|---|---|---|---|---|
| John | Director | New York | unclassified | 0.5 |
| Paul | Janitor | New York | restricted | 0.7 |
| Dave | Analyst | Paris | confidential | 0.3 |
| Ellen | Field agent | Berlin | secret | 0.2 |
| Magdalen | Double agent | Paris | top secret | 1.0 |
| Nancy | HR director | Paris | restricted | 0.8 |
| Susan | Analyst | Berlin | secret | 0.2 |

What is the probability to have a tuple with value New York?

# Example of PQE on TID

| name | position | city | classification | prob |
|---|---|---|---|---|
| John | Director | New York | unclassified | 0.5 |
| Paul | Janitor | New York | restricted | 0.7 |
| Dave | Analyst | Paris | confidential | 0.3 |
| Ellen | Field agent | Berlin | secret | 0.2 |
| Magdalen | Double agent | Paris | top secret | 1.0 |
| Nancy | HR director | Paris | restricted | 0.8 |
| Susan | Analyst | Berlin | secret | 0.2 |

What is the probability to have a tuple with value **New York**?

- It is **one minus** the probability of not having such a tuple

# Example of PQE on TID

| name | position | city | classification | prob |
|---|---|---|---|---|
| John | Director | New York | unclassified | 0.5 |
| Paul | Janitor | New York | restricted | 0.7 |
| Dave | Analyst | Paris | confidential | 0.3 |
| Ellen | Field agent | Berlin | secret | 0.2 |
| Magdalen | Double agent | Paris | top secret | 1.0 |
| Nancy | HR director | Paris | restricted | 0.8 |
| Susan | Analyst | Berlin | secret | 0.2 |

What is the probability to have a tuple with value **New York**?

- It is **one minus** the probability of not having such a tuple
- Not having such a tuple is the **independent AND** of not having each tuple

# Example of PQE on TID

| name | position | city | classification | prob |
|---|---|---|---|---|
| John | Director | New York | unclassified | 0.5 |
| Paul | Janitor | New York | restricted | 0.7 |
| Dave | Analyst | Paris | confidential | 0.3 |
| Ellen | Field agent | Berlin | secret | 0.2 |
| Magdalen | Double agent | Paris | top secret | 1.0 |
| Nancy | HR director | Paris | restricted | 0.8 |
| Susan | Analyst | Berlin | secret | 0.2 |

What is the probability to have a tuple with value New York?

- It is one minus the probability of not having such a tuple
- Not having such a tuple is the independent AND of not having each tuple
- So the result is $1 - (1 - 0.5) \times (1 - 0.7) = 0.85$

# Complexity of PQE

Formal question:

- We fix a Boolean query, e.g., $\exists xy\ R(x), S(x, y), T(y)$

# Complexity of PQE

Formal question:

- We **fix** a Boolean query, e.g., $\exists xy\ R(x), S(x,y), T(y)$
- We are given a **tuple-independent database** $D$, i.e., a relational database where facts are independent and have probabilities

# Complexity of PQE

Formal question:

- We **fix** a Boolean query, e.g., $\exists xy\ R(x), S(x,y), T(y)$
- We are given a **tuple-independent database** $D$, i.e., a relational database where facts are independent and have probabilities
- Can we **compute** the total probability of the possible worlds of $D$ that satisfy $Q$?

# Complexity of PQE

Formal question:

- We **fix** a Boolean query, e.g., $\exists xy\ R(x), S(x,y), T(y)$
- We are given a **tuple-independent database** $D$, i.e., a relational database where facts are independent and have probabilities
- Can we **compute** the total probability of the possible worlds of $D$ that satisfy $Q$?
- Note that we study **data complexity**, i.e., $Q$ is **fixed** and the input is $D$

# Naive probabilistic query evaluation

- Consider all possible worlds of the input

# Naive probabilistic query evaluation

- Consider all possible worlds of the input
- Run the query over each possible world

# Naive probabilistic query evaluation

- Consider all possible worlds of the input
- Run the query over each possible world
- Sum the probabilities of all worlds that satisfy the query

# Naive probabilistic query evaluation example

TID $D$

| in | out | |
|---|---|---|
| A | B | 0.8 |
| B | C | 0.2 |

Query $Q$

$$R(x, y) \land R(y, z)$$

# Naive probabilistic query evaluation example

TID $D$

| in | out | |
|---|---|---|
| A | B | 0.8 |
| B | C | 0.2 |

Query $Q$

$$R(x, y) \wedge R(y, z)$$

Possible worlds and probabilities:

| in | out |
|---|---|
| A | B |
| B | C |

$0.8 \times 0.2$

| in | out |
|---|---|
| A | B |
| B | C |

$(1 - 0.8) \times 0.2$

| in | out |
|---|---|
| A | B |
| B | C |

$0.8 \times (1 - 0.2)$

| in | out |
|---|---|
| A | B |
| B | C |

$(1 - 0.8) \times (1 - 0.2)$

# Naive probabilistic query evaluation example

TID $D$

| in | out | |
|---|---|---|
| A | B | 0.8 |
| B | C | 0.2 |

Query $Q$

$R(x, y) \wedge R(y, z)$

Possible worlds and probabilities:

| in | out |
|---|---|
| A | B |
| B | C |

$0.8 \times 0.2$

| in | out |
|---|---|
| A | B |
| B | C |

$(1 - 0.8) \times 0.2$

| in | out |
|---|---|
| A | B |
| B | C |

$0.8 \times (1 - 0.2)$

| in | out |
|---|---|
| A | B |
| B | C |

$(1 - 0.8) \times (1 - 0.2)$

Total probability that $Q$ holds: $0.8 \times 0.2 = 0.16$.

# Naive evaluation advantages and drawbacks

- Naive evaluation is always possible

# Naive evaluation advantages and drawbacks

- Naive evaluation is **always possible**
- However, it takes **exponential time** in general
  - $\rightarrow$ Even if the query output has **few possible worlds**!
  - $\rightarrow$ Feasible if the **input** has few possible worlds (few tuples)

## Naive evaluation advantages and drawbacks

- Naive evaluation is **always possible**
- However, it takes **exponential time** in general
  - → Even if the query output has **few possible worlds**!
  - → Feasible if the **input** has few possible worlds (few tuples)
- In fact, naive evaluation is in **#P**
  - → Can be expressed (up to normalization) as the **number of accepting paths** of a **nondeterministic PTIME Turing machine**
  - → To see why: **guess** a possible world (with the right probabilities) and **check** the query

## Naive evaluation advantages and drawbacks

- Naive evaluation is **always possible**
- However, it takes **exponential time** in general
  - → Even if the query output has **few possible worlds**!
  - → Feasible if the **input** has few possible worlds (few tuples)
- In fact, naive evaluation is in **#P**
  - → Can be expressed (up to normalization) as the **number of accepting paths** of a **nondeterministic PTIME Turing machine**
  - → To see why: **guess** a possible world (with the right probabilities) and **check** the query
- Probabilistic query evaluation is **computationally intractable** so it is unlikely that we can beat naive evaluation **in general**

## Naive evaluation advantages and drawbacks

- Naive evaluation is **always possible**
- However, it takes **exponential time** in general
  - → Even if the query output has **few possible worlds**!
  - → Feasible if the **input** has few possible worlds (few tuples)
- In fact, naive evaluation is in **#P**
  - → Can be expressed (up to normalization) as the **number of accepting paths** of a **nondeterministic PTIME Turing machine**
  - → To see why: **guess** a possible world (with the right probabilities) and **check** the query
- Probabilistic query evaluation is **computationally intractable** so it is unlikely that we can beat naive evaluation **in general**
  - → But **some queries** admit an efficient algorithm!

## Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?

# Some examples of PQE

- What is the probability of the query: $\exists x\, R(x)$?
  - It asks: "do we have an $R$-fact?"

## Some examples of PQE

- What is the probability of the query: $\exists x\, R(x)$?
  - It asks: "do we have an $R$-fact?"

$\rightarrow$ It is:

# Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"

  $\rightarrow$ It is: $1 -$

# Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"

$\rightarrow$ It is: $1 - \prod_{R(a)}$

## Some examples of PQE

- What is the probability of the query: $\exists x \, R(x)$?
  - It asks: "do we have an $R$-fact?"

$\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

## Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"

$\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?

# Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"

$\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy\ R(x), S(x,y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"

# Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"

  $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"
  - Idea: case disjunction based on the value of $x$

## Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"
  
  $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"
  - Idea: case disjunction based on the value of $x$
  - We get:

## Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "**do we have an $R$-fact?**"
  - $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: "**is there an $R$-fact which also has an $S$-fact?**"
  - Idea: **case disjunction** based on the value of $x$
  - We get: $1 -$

# Some examples of PQE

- What is the probability of the query: $\exists x \; R(x)$?
    - It asks: "do we have an $R$-fact?"
  
  $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy \; R(x), S(x, y)$?
    - It asks: "is there an $R$-fact which also has an $S$-fact?"
    - Idea: case disjunction based on the value of $x$
    - We get: $1 - \prod_{a}$

## Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"
  - $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$
- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"
  - Idea: case disjunction based on the value of $x$
  - We get: $1 - \prod_a \big(1 -$

# Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"
  
  $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"
  - Idea: case disjunction based on the value of $x$
  - We get: $1 - \prod_a \big(1 - \Pr(R(a)) \times$

## Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"
  
  $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$
- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"
  - Idea: case disjunction based on the value of $x$
  - We get: $1 - \prod_a \left(1 - \Pr(R(a)) \times \left(1 - \right.\right.$

# Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"
  - $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$
- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"
  - Idea: case disjunction based on the value of $x$
  - We get: $1 - \prod_a \big(1 - \Pr(R(a)) \times \big(1 - \prod_b$

# Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: “do we have an $R$-fact?”
  
  $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: “is there an $R$-fact which also has an $S$-fact?”
  - Idea: case disjunction based on the value of $x$
  - We get: $1 - \prod_a \left(1 - \Pr(R(a)) \times \left(1 - \prod_b (1 -\right.\right.$

## Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"

  $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"
  - Idea: case disjunction based on the value of $x$
  - We get: $1 - \prod_a \left(1 - \Pr(R(a)) \times \left(1 - \prod_b (1 - \Pr(S(a, b)))\right)\right)$

## Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"
  - $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"
  - Idea: case disjunction based on the value of $x$
  - We get: $1 - \prod_a \left(1 - \Pr(R(a)) \times \left(1 - \prod_b (1 - \Pr(S(a, b)))\right)\right)$
  - Make sure you understand why everything is independent in this case!

# Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"
  - $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$
- What is the probability of the query: $\exists xy\ R(x), S(x, y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"
  - Idea: case disjunction based on the value of $x$
  - We get: $1 - \prod_a \left(1 - \Pr(R(a)) \times \left(1 - \prod_b (1 - \Pr(S(a, b)))\right)\right)$
  - Make sure you understand why everything is independent in this case!
- What is the probability of the query: $\exists xy\ R(x), S(x, y), T(y)$?

# Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "do we have an $R$-fact?"
  - $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$
- What is the probability of the query: $\exists xy\ R(x), S(x,y)$?
  - It asks: "is there an $R$-fact which also has an $S$-fact?"
  - Idea: case disjunction based on the value of $x$
  - We get: $1 - \prod_a \left(1 - \Pr(R(a)) \times \left(1 - \prod_b (1 - \Pr(S(a,b)))\right)\right)$
  - Make sure you understand why everything is independent in this case!
- What is the probability of the query: $\exists xy\ R(x), S(x,y), T(y)$?
  - This one is #P-hard!

## Some examples of PQE

- What is the probability of the query: $\exists x\ R(x)$?
  - It asks: "**do we have an *R*-fact?**"
  
  $\rightarrow$ It is: $1 - \prod_{R(a)} (1 - \Pr(R(a)))$

- What is the probability of the query: $\exists xy\ R(x), S(x,y)$?
  - It asks: "**is there an *R*-fact which also has an *S*-fact?**"
  - Idea: **case disjunction** based on the value of $x$
  - We get: $1 - \prod_a \left(1 - \Pr(R(a)) \times \left(1 - \prod_b (1 - \Pr(S(a,b)))\right)\right)$
  - Make sure you understand **why** everything is independent in this case!

- What is the probability of the query: $\exists xy\ R(x), S(x,y), T(y)$?
  - This one is **#P-hard**!

**Research question:** can we **characterize the easy cases** and **prove hardness otherwise**?