Logic and Databases

Phokion G. Kolaitis

UC Santa Cruz & IBM Research - Almaden

Lecture 4 - Part 2

Alternative Semantics of Queries

- Bag Semantics

  We focused on the containment problem for conjunctive queries under bag semantics.

  Next, we will discuss:

- Probabilistic Databases
- Inconsistent Databases

  The focus will be on the data complexity of conjunctive queries in these two frameworks.

Probabilistic Databases

- So far, data stored in a database have been assumed to exist with certainty
- However, in modern applications, data may be uncertain: noisy, fuzzy, corrupted, or even missing.
  - Such applications include social media, information integration, scientific data management, . . .
- Probabilistic Databases provide a framework for modeling and managing uncertain data.
  - Probabilistic Databases extend relational databases with probabilities.
  - Both the data and their probabilities are stored as "standard" relations, but the semantics of query answering takes probabilities into account.

Probabilistic Databases

Definition

A probabilistic database is a pair $\mathbf{W} = (\mathbf{D}, P)$ such that

- $\mathbf{D} = \{D_1, \ldots, D_k\}$ is a finite set of databases D_i over the same schema.
- $P : \mathbf{D} \to [0, 1]$ is a function such that $\sum_{i=1}^{k} P(D_k) = 1$.

Intuition

- A probabilistic database can be in one of finitely many possible states, each with some probability.
- $\mathbf{D}$ is a set of possible worlds representing the possible states of the probabilistic database.

Marginal Probabilities

Definition

Let $\mathbf{W} = (\mathbf{D}, P)$ be a probabilistic database.

- Let q be a k-ary query, $k \geq 1$, and let $\mathbf{a}$ be a k-tuple. The marginal probability $Pr(q, \mathbf{a}, \mathbf{W})$ of $\mathbf{a}$ is

$$Pr(q, \mathbf{a}, \mathbf{W}) = \sum_{\mathbf{a} \in q(D_i)} P(D_i).$$

- Let q be a Boolean query. The marginal probability $Pr(q, \mathbf{W})$ of q is

$$Pr(q, \mathbf{W}) = \sum_{D_i \models q} P(D_i).$$

Query Evaluation over Probabilistic Databases

- Query Evaluation over probabilistic databases:
  Given a k-ary query q, a k-tuple $\mathbf{a}$, and a probabilistic database $\mathbf{W}$, compute the marginal probability $Pr(q, \mathbf{a}, \mathbf{W})$.
- Note that this is a combined complexity problem.
  Here, we will focus on the data complexity of Boolean conjunctive queries over probabilistic databases.
- Fix a Boolean conjunctive query q.
  Then $Pr[q]$ is the following algorithmic problem:
  Given a probabilistic database $\mathbf{W}$, compute the marginal probability $Pr(q, \mathbf{W})$.

Representations of Probabilistic Databases

- A probabilistic database may have an arbitrarily large number of possible worlds, which implies that listing all these possible worlds may be infeasible.
- For this reason, several different compact representations of probabilistic databases have been introduced and investigated.
- Here, we will focus on tuple-independent databases, which is arguably the simplest model for probabilistic database design.
- Intuitively, in a tuple-independent database all tuples are independent probabilistic events.

Tuple-Independent Databases

- A tuple-independent relation is a relation $R(A_1, \ldots, A_m, P)$ in which tuples $(a_1, \ldots, a_m)$ are independent events and the values of P are numbers in the interval $[0, 1]$ denoting the marginal tuple probabilities of the tuples.

Company	Product	P
Apple	iphone 6	0.95
Samsung	Galaxy 7	0.96
Apple	iphone 7	0.75
Microsoft	Lumia 640	0.85

  - This table is a compact representation of 16 possible tables.
  - For example, the table consisting of the first, the second, and the fourth tuple has probability $0.95 \cdot 0.96 \cdot 0.25 \cdot 0.85$.
- A tuple-independent database is a database consisting of tuple-independent relations.

Data Complexity in Tuple-Independent Databases

- Fix a Boolean query q.
  $Pr[q]$ is the following problem: Given a tuple-independent database $\mathbf{W}$, compute the marginal probability $Pr(q, \mathbf{W})$.
- This is a data complexity problem because the query is fixed and the input is a tuple-independent database $\mathbf{W}$.
- Recall that the data complexity of unions of conjunctive queries on (deterministic) databases is in LOGSPACE.

Data Complexity in Tuple-Independent Databases

Dichotomy Theorem (Dalvi and Suciu - 2012)
If q is a union of Boolean conjunctive queries, then $Pr[q]$ is in P or $Pr[q]$ is $\#\mathrm{P}$-complete.

Data Complexity in Tuple-Independent Databases

Dichotomy Theorem (Dalvi and Suciu - 2012)

If q is a union of Boolean conjunctive queries, then $Pr[q]$ is in P or $Pr[q]$ is $\#\mathrm{P}$-complete.

Note

- $\#\mathrm{P}$ is the class of counting problems associated with decision problems in NP.
- The prototypical $\#\mathrm{P}$-complete problem is $\#$SAT:
  Given a CNF-formula φ, compute the number of its satisfying assignments.
- Valiant (1979) also showed that $\#$POSITIVE 2SAT is $\#\mathrm{P}$-complete.
  ("easy" decision - "hard" counting phenomenon)

Hierarchical Queries

Definition

- A self-join free conjunctive query is a conjunctive query in which no relation symbol appears more than once.
- Let q be a self-join free conjunctive query.
  - If x is a variable of q, then $at(x)$ is the set of all atoms of x in which x appears.
  - We say that q is hierarchical if for every two variables x and y of q, one of the following holds:
    $at(x) \subseteq at(y), \quad at(y) \subseteq at(x), \quad at(x) \cap at(y) = \emptyset.$

Example

- The query $\exists x \exists y (R(x) \wedge S(x, y))$ is hierarchical.
- The query $\exists x \exists y (R(x) \wedge S(x, y) \wedge T(y))$ is not hierarchical.

Data Complexity in Tuple-Independent Databases

The Little Dichotomy Theorem (Dalvi and Suciu - 2004)
Let q be a Boolean self-join free conjunctive query.

- If q is hierarchical, then $Pr[q]$ is in P.
- If q is not hierarchical, then $Pr[q]$ is $\#\mathrm{P}$-complete.

Data Complexity in Tuple-Independent Databases

The Little Dichotomy Theorem (Dalvi and Suciu - 2004)

Let q be a Boolean self-join free conjunctive query.

- If q is hierarchical, then $Pr[q]$ is in P.
- If q is not hierarchical, then $Pr[q]$ is $\#\mathrm{P}$-complete.

Proof Idea

- Hierarchical queries admit safe evaluation plans.
- Non-hierarchical queries:
  - Show that $Pr[\exists x \exists y (R(x) \wedge S(x, y) \wedge T(y))]$ is $\#\mathrm{P}$-complete.
  - Show that if q is not hierarchical, then $Pr[\exists x \exists y (R(x) \wedge S(x, y) \wedge T(y))]$ is reducible to $Pr[q]$.

Hierarchical Queries

Let q be the hierarchical query $\exists x \exists y (R(x) \wedge S(x, y))$ and let $\mathbf{W}$ be a tuple-independent database.

- First, write q as $\exists x (R(x) \wedge \exists y S(x, y))$.
- Then, using tuple-independence repeatedly, we have that:

$$\begin{aligned} Pr[q] &= 1 - \prod_{a \in adom(\mathbf{W})} (1 - P((R(a) \wedge \exists y S(a, y)))) \\ &= 1 - \prod_{a \in adom(\mathbf{W})} (1 - P((R(a)) \cdot P(\exists y S(a, y)))) \\ &= 1 - \prod_{a \in adom(\mathbf{W})} (1 - P((R(a)) \cdot (1 - \prod_{b \in adom(\mathbf{W})} (1 - P(S(a, b)))))) \end{aligned}$$

- The last expression has size $O(n^2)$, where $n = |adom(\mathbf{W})|$.

Non-Hierarchical Queries

- A positive partitioned 2DNF formula (PP2DNF) is a DNF-formula of the form
$$x_{i_1} y_{j_1} \vee \cdots \vee x_{i_k} y_{j_k},$$
where the x_i's and the y_j's form disjoint sets of variables.

- Theorem (Provan and Ball - 1982)
$\#$PP2DNF is $\#\mathrm{P}$-complete.

- Theorem (Dalvi and Suciu - 2004)
There is a counting reduction from $\#$PP2DNF to $Pr[\exists x \exists y (R(x) \wedge S(x, y) \wedge T(y))]$.

Non-Hierarchical Queries

Counting reduction from

$\#$PP2DNF to $Pr[\exists x \exists y (R(x) \wedge S(x, y) \wedge T(y))]$.

- Suppose φ is the formula $x_1 y_1 \vee x_1 y_2 \vee x_2 y_1$
- Let $\mathbf{W}_\varphi$ be the tuple-independence database

R	X	P	S	X	Y	P	T	Y	P
	x_1	0.5		x_1	y_1	1		y_1	0.5
	x_2	0.5		x_1	y_2	1		y_2	0.5
				x_2	y_1	1			

- There is a 1-1 correspondence between truth assignments for φ and possible worlds for $\mathbf{W}_\varphi$.
- It is easy to see that
  $$\#\varphi = 2^n Pr(\exists x \exists y (R(x) \wedge S(x, y) \wedge T(y)), \mathbf{W}_\varphi),$$
  where n is the number of variables of φ.

Non-Hierarchical Queries

Let q be a Boolean conjunctive query that is not hierarchical.

- By definition, there are variables x and y of q such that
  $$at(x) \not\subseteq at(y),\ at(y) \not\subseteq at(x),\ at(x) \cap at(y) \neq \emptyset.$$
- Since $at(x) \not\subseteq at(y)$, there is an atom $R'(x, \ldots)$ in which y does not appear.
- Since $at(y) \not\subseteq at(x)$, there is an atom $T'(y, \ldots)$ in which x does not appear.
- Since $at(x) \cap at(y) \neq \emptyset$, there is an atom $T'(x, y, \ldots)$ in which both x and y appear.
- These atoms can be used to obtain a counting reduction from $Pr[\exists x \exists y (R(x) \wedge S(x, y) \wedge T(y))]$ to $Pr[q]$.

Data Complexity in Tuple-Independent Databases

The Little Dichotomy Theorem (Dalvi and Suciu - 2004)

Let q be a Boolean self-join free conjunctive query.

- If q is hierarchical, then $Pr[q]$ is in P.
- If q is not hierarchical, then $Pr[q]$ is $\#\mathrm{P}$-complete.

Data Complexity in Tuple-Independent Databases

The Little Dichotomy Theorem (Dalvi and Suciu - 2004)
Let q be a Boolean self-join free conjunctive query.

- If q is hierarchical, then $Pr[q]$ is in P.
- If q is not hierarchical, then $Pr[q]$ is $\#\mathrm{P}$-complete.

Open Problems:

- Dichotomy Theorem for arbitrary conjunctive queries on the block-independent-disjoint model.
  - Dichotomy known for self-join free conjunctive queries.
- Dichotomy Theorem for arbitrary conjunctive queries on the tuple-independent model in the presence of functional dependencies.
  - Dichotomy known for self-join free conjunctive queries.