1 General Introduction

This course could be seen with three different angles. First, as an introduction to the theory of databases, and more specifically to the study of query languages for database systems. We will give the basis of what the complexity of a query language is. We will also give elementary tools for studying the expressive power of query languages. The second point of view is an introduction to descriptive complexity. We will see that the difficulty of expressing a problem is closely related to the resources necessary for computing it. We will see that many of the open problems concerning the separation of complexity classes can be reformulated in term of equality of logics, without any reference to a computing model such as Turing machines. The third point of view is an introduction to finite model theory. We will present concepts like locality, 0-1 laws and Ehrenfeucht-Fraïssé games. The overall objective of this course is to show that all this form a unique story linking nicely logic, complexity and query languages.

2 Signatures and finite models

A schema or signature $\sigma = \{R, F, ar\}$ consists of a set $R$ of relational symbols, a set $F$ of functional symbols and an arity function $ar$, mapping $R \cup F$ to the set of nonnegative integers. The function $ar$ gives for each relational and functional symbol the number of arguments that it takes. If $c \in F$ and $ar(F) = 0$ then we say that $c$ is a constant.
Assuming a finite universe \( \mathcal{U} \), a finite model \( M \) over a schema \( \sigma \) assigns a relation \( M_R \subset \mathcal{U}^{\text{ar}(R)} \) for any \( R \in \mathcal{R} \) and a function \( M_f : \mathcal{U}^{\text{ar}(f)} \to \mathcal{U} \) for any \( f \in F \). For any constant \( c \) a model \( M \) assigns an element \( M_c \in \mathcal{U} \). We have to stress the fact that we do not distinguish between two isomorphic models.

**Example 2.1.** Consider the graph \( G \) with the vertex set \( V(G) = \{a, b, c, d\} \) and the edge set \( E(G) = \{(c, d), (c, b), (b, a), (a, c)\} \). Then, this graph can be seen as a model \( M \) over the signature \( \sigma = \{\mathcal{E}, \emptyset, \text{ar}\} \) with \( M_E = E(G) \) and \( \mathcal{U} = V(G) \). As stated above, we could have the model \( M' \) using the universe \( \mathcal{U}' = \{1, 2, 3, 4\} \), with \( M'_E = \{(3, 4), (3, 2), (2, 1), (1, 3)\} \). Since \( M' \) is obtained from \( M \) by using an isomorphic domain, we will consider \( M = M' \).

**Example 2.2.** Any binary word \( w \in \{a, b\}^* \) can be seen as a model \( M \) over the signature \( \sigma = \{\{P_a, P_b, <\}, \emptyset, \text{ar}\} \). Thus, if \( w = abaaba \) then we have \( M_{P_a} = \{1, 3, 4, 6\} \), \( M_{P_b} = \{2, 5\} \), and the universe \( \mathcal{U} = \{1, 2, 3, 4, 5, 6\} \). The relation \( M_\prec \) is the natural order relation over \( \mathcal{U} \).

## 3 Query languages

### 3.1 First Order Logic

**Definition 3.1** (first order logic). Let \( \sigma = (\mathcal{R}, \mathcal{F}, \text{ar}) \) be a signature. Also, consider \( V = \{x, y, z, \cdots\} \) a fixed, countable set of variables which range over the universe which gives semantics to the expression.

As follows, we inductively define the set of terms over \( \sigma \). First, let any variable \( x \in V \) be a term. Then, let \( f(t_1, \cdots, t_k) \) be a term, where \( f \) is a functional symbol with \( \text{ar}(f) = k \), and \( t_1, \cdots, t_k \) are terms. Notice that by taking \( k = 0 \) in the above definition it follows that any constant is also a term.

First order expressions are defined inductively as well. Let \( R(t_1, \cdots, t_k) \) be an atomic first order expression, where \( R \) is a relational symbol with \( \text{ar}(R) = k \) and \( t_1, \cdots, t_k \) are terms. If \( \phi \) and \( \psi \) are first order expressions, then \( \neg \phi, (\phi \lor \psi), (\phi \land \psi) \) are first order expressions as well. Finally, if \( \phi \) is a first order expression and \( x \) is a variable then \( (\exists x \phi) \) and \( (\forall x \phi) \) are first order expressions as well.

Given a first order expression, its free variables are the variables not within the scope of any quantifier. If a formula \( \psi \) has the set of free variables \( x \) then we may denote \( \psi \) by \( \psi(x) \). A formula with no free variables is called a sentence.
Let $\phi$ be a first order expression over the fixed vocabulary $\sigma$, and $M$ be an appropriate model over $\sigma$. As follows, we will define when $M$ satisfies $\phi$, denoted as $M \models \phi$. If $\phi$ contains free variable, $\equiv$ assumes a function $\nu$ assigning elements of $M$ to the free variables of $\phi$.

First, notice that $\nu$ can be extended to terms in the obvious way: Inductively, if $t = x$ is a variable, then $\nu(t) = \nu(x)$. Otherwise, if $t = f(t_1, \ldots, t_k)$, where $f$ is a functional symbol with $\text{ar}(f) = k$ and $t_1, \ldots, t_k$ are terms, we have $\nu(t) = f_M(\nu(t_1), \ldots, \nu(t_k))$.

Now suppose $\phi$ is a first order expression. We define satisfaction on $\phi$ using structural induction. If $\phi$ is an atomic expression, $\phi = R(t_1, \ldots, t_k)$, then $M \models \phi$ iff $(\nu(t_1), \ldots, \nu(t_M)) \in R_M$. If $\phi = \neg \psi$, then $M \models \phi$ iff $M \not\models \psi$. If $\phi = \psi_1 \lor \psi_2$ or $\phi = \psi_1 \land \psi_2$ then $M \models \phi$ iff $M \models \psi_1$ or $M \models \psi_2$, respectively. $M \models \psi_1$ and $M \models \psi_2$. If $\phi = (\exists x \psi)$ then $M \models \phi$ iff there exists $u \in U$ such that $M \models \psi$ after extending $\nu$ with $\nu(x) = u$.

If $\psi$ is a sentence, then we denote by $L_\psi = \{M | M \models \psi\}$ the language of $\psi$.

**Example 3.1.** Consider the signature for words $\sigma = \{\{P_a, P_b, <\}, \emptyset, \text{ar}\}$. Then, the sentence $\psi = \forall y(P_b(y) \rightarrow \exists x(x < y \land P_a(x) \land \forall z(\neg(x < z < y))) \land \forall x(P_a(x) \rightarrow \exists y(x < y \land P_a(x) \land \forall z(\neg(x < z < y)))$ has $L(\psi) = (ab)^*$.

**Example 3.2.** Consider the signature for graphs $\sigma = \{\{E\}, \emptyset, \text{ar}\}$. Then, the sentence $\psi = \forall x \exists y_1 \exists y_2 \exists y_3 \forall y(E(x, y) \rightarrow (y = y_1 \lor y = y_2 \lor y = y_3))$ is satisfied by all models over $\sigma$ which correspond to graphs where every node has degree less than or equal to 3.

As follows, we will give decidability results for first order satisfiability and equivalence problems. The satisfiability problem for $FO$, denoted as $\text{SAT}(FO)$ asks the question of whether a $FO$ sentence is satisfiable. Thus, the input is any $FO$ sentence over a signature $\sigma$ and the output is ACCEPT if there exists a model $M$ over $\sigma$ such that $M \models \psi$ and REJECT otherwise. The first order equivalence problem $\text{EQ}(FO)$ takes as input two first order sentences $\phi$ and $\psi$. It outputs ACCEPT when $\phi \leftrightarrow \psi$ and REJECT otherwise.

**Theorem 3.1.** The problem $\text{SAT}(FO)$ is undecidable.

*Proof.* The proof can be found in [1].

**Theorem 3.2.** The problem $\text{EQ}(FO)$ is undecidable.

*Proof.* This can be proved by reducing $\text{SAT}(FO)$ to $\text{EQ}(FO)$. Indeed, having a formula $\psi$ we can check if it is unsatisfiable by testing equivalence between $\psi$ and $false$, where $false$ denotes any unsatisfiable formula.
3.2 Conjunctive queries

**Definition 3.2 (CQ).** Let $Q$ be an FO formula over a signature $\sigma$. Then, $Q$ is a conjunctive query (CQ) if it is a conjunction of atoms over free and existentially quantified variables. Thus, it has the form $(\exists \bar{y}(R_1(x_1) \land \cdots \land R_k(x_k)))$, where $\bar{y} \subseteq (x_1 \cup \cdots \cup x_k)$. The variables in $\bar{y}$ are called bounded, while the others are called free. Assuming that $\bar{z}$ is the set of free variables, we have that $\text{vars}(Q) = \bar{y} \cup \bar{z}$, where $\text{vars}(Q)$ denotes the set of variables in $Q$. We can rewrite $Q$ in a more convenient manner as

$$Q(\bar{z}) \leftarrow R_1(x_1), \ldots, R_k(x_k),$$

where $\bar{z}$ is called the head of the query, and the right part is called the body.

Finally, given a CQ $Q$, let $M_Q$ denote the model over $\sigma$ induced by the body of $Q$. More precisely this is the model whose universe is the set of variables of $Q$ and whose tuples are defined by $(x_1, \ldots, x_k) \in R_{M_Q}$ iff $R(x_1, \ldots, x_k)$ is an atom in $Q$.

**Example 3.3.** Consider the signature for graphs $\sigma = \{E, 0, ar\}$. Then, the CQ query $Q \leftarrow E(x, y), E(y, z), E(z, x)$ is satisfied by all models corresponding to graphs that have a triangle.

**Theorem 3.3.** Let $\sigma$ be a signature and $\varphi(\bar{x}) \leftarrow R_1(\bar{u}_1), \ldots, R_k(\bar{w}_k)$ be a CQ over $\sigma$. Let $M$ be a model over $\sigma$. Then, $M \models \varphi(\bar{a})$ iff there exists a homomorphism $h : M_\varphi \rightarrow M$ such that $h(\bar{x}) = \bar{a}$.

**Proof.** Let $\bar{y}$ be the set of bounded variables in $\varphi(\bar{x})$. Now suppose that $M \models \varphi(\bar{a})$. It follows that for any variable $z \in \bar{y}$ there exists $f(z) \in U$ such that $\varphi(\bar{a})$ is validated. As follows, consider $h : M_\varphi \rightarrow M$ defined as

$$h(z) = \begin{cases} a_i, & \text{if } z = x_i \\ f(z), & \text{otherwise} \end{cases}$$

Indeed, $h$ is a homomorphism which satisfies $h(\bar{x}) = \bar{a}$.

It remains to show that if there exists a homomorphism $h : M_\varphi \rightarrow M$ such that $h(\bar{x}) = \bar{a}$, then $M \models \varphi(\bar{a})$. From the hypothesis it follows that $h$ is a assignment of the bounded variables of $\varphi$ such that $R_i(h(x_i))$ is satisfied for all $1 \leq i \leq k$, which concludes $M \models \varphi(\bar{a})$.

**Theorem 3.4.** Let $\sigma$ be a signature and $\varphi(\bar{x}), \psi(\bar{x})$ be two CQs over $\sigma$. We say that $\varphi(\bar{x}) \subseteq \psi(\bar{x})$ iff for any model $M$, and any tuple $\bar{a}$, $M \models \varphi(\bar{a})$ implies $M \models \psi(\bar{a})$. Then, we have that $\varphi(\bar{x}) \subseteq \psi(\bar{x})$ iff there exists a homomorphism $h : M_\psi \rightarrow M_\varphi$ which satisfies $h(\bar{x}) = \bar{x}$.
Proof. Suppose that there exists a homomorphism $h : M_\psi \rightarrow M_\varphi$ which satisfies $h(\bar{x}) = \bar{x}$. Then we will show that $\varphi(\bar{x}) \subseteq \psi(\bar{x})$. Let $M$ be a model such that $M \models \varphi(\bar{a})$. Using theorem 3.3 we have that there exists a homomorphism $h' : M_\varphi \rightarrow M$ such that $h'(\bar{x}) = \bar{a}$. Thus, we have the homomorphism $(h' \circ h) : M_\psi \rightarrow M$ which, using theorem 3.3, implies that $M \models \psi(\bar{a})$.

Finally, supposing that $\varphi(\bar{x}) \subseteq \psi(\bar{x})$ we will show that there exists a homomorphism $h : M_\psi \rightarrow M_\varphi$ which satisfies $h(\bar{x}) = \bar{x}$. This follows from the fact that $M_\varphi \models \varphi(\bar{x})$. Thus, we have that $M_\varphi \models \psi(\bar{x})$ which, using theorem 3.3, implies that there exists a homomorphism $h : M_\psi \rightarrow M_\varphi$, which concludes the proof.

It is easy to see that all CQ formulas are satisfiable. Indeed, for any CQ $\psi$ we have that $M_\psi \models \psi$. As follows we define the inclusion problem for CQs, denoted $\text{INC}(CQ)$. As input, we take two CQ sentences $\varphi, \psi$. We output ACCEPT if they $\varphi$ is included into $\psi$ and REJECT otherwise.

Theorem 3.5. $\text{INC}(CQ)$ is NP-complete.

Proof. We define the problem $\text{HOM}(G_1, G_2)$ as follows: we accept if there exists a homomorphism $h : G_1 \rightarrow G_2$ and reject otherwise. Using theorem 3.6 it follows that $\text{HOM}(G_1, G_2) \leq_p \text{INC}(CQ)$. It remains now to show that $\text{HOM}(G_1, G_2)$ is NP-complete. Indeed, $\text{HOM}(G_1, G_2)$ is in NP because we can guess the homomorphism and then check if it is indeed a homomorphism. The NP-hardness part can be proved by reducing 3-COLORABILITY to $\text{HOM}(G_1, G_2)$, by observing that a graph $G$ is 3-colorable iff there exists a homomorphism from $G$ to the graph $\{(1, 2), (2, 3), (3, 1)\}$.

4 Exercises

Exercise 1: $\text{EQ}(CQ)$ is the equivalence problem for CQ. Show that $\text{EQ}(CQ)$ is NP-complete.

Exercise 2: Show that over words, the language $(aa)^*$ is not expressible in FO.

Exercise 3: Show that over graphs, connectivity is not expressible in FO.

References