First-Order Query Expressiveness

Recall:
- Syntax: a query expression \( q \) is a word from a query language (algebra expression, logical expression, etc.)
- Semantics: a query mapping \( M \) is a function that maps a database instance \( J \) to a database table \( M(J) \)
- We only study generic queries, which are closed under bijective renaming (isomorphism of databases)

**Definition 11.1:** The expressiveness of a query language is characterised by the set of query mappings that it can express. Given a query language \( L \), a query mapping \( M \) is \( L \)-definable if there is a query expression \( q \in L \) such that \( M[q] = M \).

We can study expressiveness for all query mappings over all possible databases, or we can restrict attention to a subset of query mappings or to a subset of databases.
Boolean Query Mappings

A **Boolean query mapping** is a query mapping that returns “true” (usually a database with one table with one empty row) or “false” (usually an empty database).

Every Boolean query mapping
- defines set of databases for which it is true
- defines a decision problem over the set of all databases
- could be decidable or undecidable
- if decidable, it may be characterised in terms of complexity

**Note:** the “complexity of a mapping” is always “data complexity,” i.e., complexity w.r.t. the size of the input database; the mapping defines the decision problem and is fixed.

---

The Limits of FO Queries

**Are there polynomial query mappings that cannot be expressed in FO?**

~→ **yes!**

We already knew this from previous lectures:
- We learned that $\text{AC}^0 \subseteq \text{NC}^1 \subseteq \cdots \subseteq P$
- Hence, there is a problem $X$ in $\text{NC}^1$ that is not in $\text{AC}^0$
- Therefore, the corresponding query mapping $M_X$ is not FO-definable

$\text{AC}^0 \subseteq \text{NC}^1$ was first shown for the problem $X = \text{Parity}$:
- **Input:** finite relational structure $I$
- **Output:** “true” if $I$ has an even number of domain elements

The original proof is specific to this problem [Ajtai 1983].

---

Any Other FO-Undefinable Problems?

**Yes, many.**

Strong evidence from complexity theory:
- If any P-complete problem $X$ were FO-definable,
  - then every problem in $P$ could be LogSpace-reduced to $X$
  - and then solved in $\text{AC}^0$,
  - hence every problem in $P$ could be solved in LogSpace,
  - that is, $P = \text{L}$.
- Most experts do not think that this is the case.

Therefore, one would expect all P-hard and similarly all NL-hard problems to not be FO-definable.

~→ **How can we see this more directly?**
Proving FO-Undefinability

How to show that a query mapping is FO-definable?
~ Find an FO query that expresses the query mapping

How to show that a query mapping is not FO-definable?
~ Not so easy . . . important tools:
  • Ehrenfeucht-Fraïssé games
  • Locality theorems

Ehrenfeucht-Fraïssé Games

A method for showing that certain finite structures cannot be distinguished by certain FO formulas

General idea:
• A game is played on two databases \( I \) and \( J \)
• There are two players: the Spoiler and the Duplicator
• The players select elements from \( I \) and \( J \) in each round
• Spoiler wants to show that the two databases are different
• Duplicator wants the databases to appear to be the same

We will always play on finite structures without constant symbols
(remember that one can simulate constants by unary relations with one row)

Playing One Run of an EF Game

A single run of the game has a fixed number \( r \) of rounds

Spoiler starts each round, and Duplicator answers:
• Spoiler picks a domain element from \( I \) or from \( J \)
• Duplicator picks an element from the other database (\( J \) or \( I \))

~ One element gets picked from each \( I \) and \( J \) per round

~ Run of game ends with two lists of elements:
  \( a_1, \ldots, a_r \in \Delta^I \) and \( b_1, \ldots, b_r \in \Delta^J \)

Duplicator wins the run if:
• For all indices \( i \) and \( j \), we have \( a_i = a_j \) if and only if \( b_i = b_j \).
• For all lists of indices \( i_1, \ldots, i_k \) and \( n \)-ary relation names \( R \),
  we have \( (a_{i_1}, \ldots, a_{i_k}) \in R^I \) if and only if \( (b_{i_1}, \ldots, b_{i_k}) \in R^J \).
  "The substructures induced by the selected elements are isomorphic" 

Otherwise Spoiler wins the run.

Example: Run of a Two-Turn EF Game

• edges denote a bi-directional binary predicate
• all edges are the same predicate
Example: Run of a Three-Turn EF Game

- edges denote a bi-directional binary predicate
- all edges are the same predicate

Winning the EF Game

The game is won by whoever has a winning strategy:

A player has a winning strategy if he/she can make sure that he/she will win, whatever the other player is doing.

In other words:

- Duplicator wins if he can duplicate any move that the spoiler makes.
- Spoiler wins if she can spoil any attempt to duplicate her moves.

We write $I \sim r J$ if Duplicator wins the $r$-round EF game on $I$ and $J$.

Observation: given enough moves, the spoiler will always win, unless the structures are isomorphic

Example

Who wins the 2-round game?
Who wins the 3-round game?

- edges denote a bi-directional binary predicate
- all edges are the same predicate

Quantifier Rank

EF games characterise expressivity of FO formulae based on the nesting depth of quantifiers:

- A formula without quantifiers has quantifier rank 0
- $\exists x (C(x) \land \forall y (R(x, y) \rightarrow x = y) \land \exists v S(x, v))$ has quantifier rank 2

Definition 11.4: We write $I \equiv_r J$ if $I$ and $J$ satisfy the same FO sentences of rank $r$ (or less).

Markus Krötzsch, 15th May 2019
Database Theory
slide 13 of 38

Markus Krötzsch, 15th May 2019
Database Theory
slide 14 of 38

Markus Krötzsch, 15th May 2019
Database Theory
slide 15 of 38

Markus Krötzsch, 15th May 2019
Database Theory
slide 16 of 38
The idea can be generalised to formulae \( \phi \) that are not in prenex normal form (by interleaving the choice of the quantifier and the evaluation of the formula).

**Proof idea (2)**

We outline the proof for the direction that is more important to us:

**Lemma 11.6:** For every \( r \), we find \( \models \varphi \iff \models \neg \varphi \).

**Proof (continued):** Therefore, by (+), after \( r \) rounds we have selected elements \( a_1, \ldots, a_i \in \Delta^I \) and \( b_1, \ldots, b_i \in \Delta^J \), such that \( J, [x_1 \mapsto a_1, \ldots, x_i \mapsto a_i] \models \varphi \) and \( J', [x_1 \mapsto b_1, \ldots, x_i \mapsto b_i] \not\models \varphi \).

Hence, the substructures induced by the selected elements are not isomorphic (if they were, we would find that \( \psi \) evaluates to the same in both cases).

\[ \neg \models \psi \]

Spoiler wins

The idea can be generalised to formulae \( \phi \), that are not in prenex normal form (by interleaving the choice of the quantifier and the evaluation of the formula).
Using EF Games to Show FO-Undefinability

How to show that a query mapping $M$ can not be FO-defined:

- Let $C_M$ be the class of all databases recognised by $M$
- Find sequences of databases $I_1, J_2, J_3, \ldots \in C_M$ and databases $J_1, J_2, J_3, \ldots \notin C_M$, such that $I_i \sim J_i$

$\sim$ for any formula $\varphi$ (however large its quantifier rank $r$), there is a counterexample $I_r \in C_M$ and $J_r \notin C_M$ that $\varphi$ cannot distinguish

Problems:

- How to find such sequences of $I_i$ and $J_i$?
- No general strategy exists
- Given suitable sequences, how to show that $I_i \sim J_i$?
- Can be difficult, but doable for some special cases

Expressiveness on Linear Orders

Let’s look at some very simple structures:

**Definition 11.7:** A structure $I$ is a linear order if it has a single binary predicate $\preceq$ interpreted as a total, transitive, reflexive and asymmetric relation.

**Example 11.8:** Consider the following structures:

$L_6: 1 \preceq 2 \preceq 3 \preceq 4 \preceq 5 \preceq 6$

$L_7: 1 \preceq 2 \preceq 3 \preceq 4 \preceq 5 \preceq 6 \preceq 7$

Spoiler can win the 3-round EF game as follows:

- Spoiler plays $4$ in $L_7$
- Duplicator plays $4$ in $L_6$: Spoiler plays $6$ in $L_7$ and wins
- Duplicator plays $6$ in $L_6$: Spoiler plays $7$ in $L_7$ and wins
- Duplicator plays $3$ in $L_6$: symmetric game (flipped horizontally)

EF Games and Linear Orders

**Theorem 11.10:** The following are equivalent:

- $L_m \sim L_n$
- either (1) $m = n$, or (2) $m \geq 2^r - 1$ and $n \geq 2^r - 1$

**Proof:**

- First, we show the contrapositive of the $\implies$ direction of the theorem:
  - If $m \neq n$ and $m \leq 2^r - 1$, then $L_m \not\sim L_n$
  - If $m \neq n$ and $n \leq 2^r - 1$, then $L_m \not\sim L_n$ (analogous to the previous case)
- We define a winning strategy for the spoiler to show the implication in (2.1) assuming that $n \geq 2^r - 1$ (other cases need to be considered separately):
EF Games and Linear Orders

Proof: We define a winning strategy for the spoiler to show the implication in (2.1) assuming that $n \geq 2^2 - 1$ (other cases need to be considered separately):

- Throughout this strategy, the spoiler always selects elements from $L_m$. Therefore, the duplicator can only select elements from $L_m$.
- For all $i \in \{1, \ldots, r\}$, let $LD_i = j - k$ and $RD_i = \ell - j$ such that:
  - The duplicator chooses the $j$-th element in the in the $i$-th move.
  - If no element to the left of the $j$-th element in $L_m$ has been picked by the duplicator, then $k = 0$. Otherwise, $k$ is the largest number such that $k \leq j$ and the $k$-th element has been selected by the duplicator.
  - If no element to the right of the $j$-th element in $L_m$ has been picked by the duplicator, then $\ell = m$. Otherwise, $\ell$ is the smallest number such that $\ell \geq j$ and the $\ell$-th element has been selected by the duplicator.
- Game strategy for the spoiler:
  - First, the spoiler picks the “middle” element in $L_r$.
  - For each turn $i \in \{2, \ldots, r\}$: if $LD_{i-1} \leq RD_{i-1}$, then the spoiler picks an element to the left of its last choice such that there are the “same” number of consecutive unselected elements to its right than to its left. Otherwise, it picks an element to the right of its last choice that satisfies the same condition.

Parity

Theorem 11.12: Parity is not FO-definable for linear orders, hence it is not FO-definable for arbitrary databases.

Proof:
- Suppose for a contradiction that Parity is FO-definable by some query $\varphi$.
- Let $r$ be the quantifier rank of $\varphi$.
- Consider databases $L_m$ and $L_n$ with $m = 2^r$ and $n = 2^r + 1$.
- We know that $L_m \not\models \varphi$, and therefore $L_m \not\equiv L_n$.
- Hence, $L_m \models \varphi$ if and only if $L_n \models \varphi$.
- But $L_m \in \text{Parity}$ while $L_n \notin \text{Parity}$.
- Therefore, $\varphi$ does not FO-define Parity. Contradiction.

Connectivity

The Connectivity problem over finite graphs is as follows:

Connectivity
- Input: A finite graph (relational structure with one binary relation “edge”)
- Output: “true” if there is an (undirected) path between any pair of vertices

Theorem 11.13: Connectivity is not FO-definable.

Proof:
- Suppose for a contradiction that Connectivity is FO-definable using a query $\varphi$.
- We show that this would make Parity FO-definable on linear orders.
- For a linear order $L$ with order predicate $\leq$, we define a finite graph $\overline{G}(L)$ over a binary predicate “edge” such that $\overline{G}(L)$ is connected if and only if $L$ has an odd number of elements.
Defining a Graph From a Linear Order

We use abbreviations for the following FO formulas:

\[
\text{succ}\{x, y\} = (x \leq y) \land \neg(y \leq x) \land \\
\forall z.(z \leq x \lor y \leq z)
\]

\[
\text{min}\{x\} = \forall x.z \leq z
\]

\[
\text{max}\{x\} = \forall z.z \leq x
\]

\[
\text{succ}^c\{x, y\} = \text{succ}\{x, y\} \lor (\text{max}\{x\} \land \text{min}\{y\})
\]

y is the successor of x

x is the first element

x is the last element

Circular version of \(\text{succ}\)

We now define the formula \(\varphi\) that derives edges from a linear order:

\[
\forall x, y.\text{edge}(x, y) \leftrightarrow \exists z.\text{succ}^c\{x, z\} \land \text{succ}^c\{z, y\}
\]

Completing the Proof

Observation:
The graph \(G(L)\) is connected if and only if \(L\) has odd parity.

Therefore, if \(\varphi\) FO-defines Connectivity on graphs with predicate edge, then \(\neg(\varphi \land \psi)\) FO-defines Parity on linear orders.

Since Parity is not FO-definable, no such \(\varphi\) can exist.

Beyond Linear Orders: Locality

Intuition: Duplicator can win an EF game if selected nodes have the same "neighbourhood"

\[
\text{\textit{let's define this for graphs (structures with binary predicates)}}
\]

Definition 11.14: Consider a graph \(G\). For a natural number \(d \geq 0\) and a vertex \(v\), the \(d\)-neighbourhood of \(v\), \(N(v, d)\), is defined inductively:

- \(N(v, 0) = \{v\}\)
- \(N(v, d + 1) = N(v, d) \cup \{w \mid w \text{ is a direct neighbour of some } w' \in N(v, d)\}\)

Two vertices \(v\) and \(w\) have the same \(d\)-type if the subgraphs \(G_{N(v, d)}\) and \(G_{N(w, d)}\) are isomorphic.

Two graphs are \(d\)-equivalent if, for every \(d\)-type, they have the same number of \(d\)-neighbourhoods of this type.
Locality and FO-definability

A special case of Gaifman’s Locality Theorem of first-order logic:

**Theorem 11.15:** For every integer $r \geq 1$:
- if $G_1$ is $3^r$-equivalent to $G_2$,
- then $G_1 \sim_r G_2$, and thus $G_1 \equiv_r G_2$

~ Intuition: FO can only express local properties

How to show that a query mapping $M$ cannot be FO-defined:
- Let $C_M$ be the class of all databases recognised by $M$
- Find sequences of graphs $I_1, I_2, I_3, \ldots \in C_M$ and graphs $J_1, J_2, J_3, \ldots \notin C_M$, such that $I_i$ is $i$-equivalent to $J_i$

~ For any formula $\varphi$ (however large its quantifier rank $r$), there is a counterexample $I_{\varphi^{-1}} \in C_M$ and $J_{\varphi^{-1}} \notin C_M$ that $\varphi$ cannot distinguish

---

**Connectivity is not FO-definable (Proof 2)**

**Theorem 11.16:** Connectivity is not FO-definable.

**Proof:** counterexample for quantifier rank $r$: set $d = 3^r$

\[
\begin{align*}
I_d & \quad J_d \\
6d & \quad 3d \\
3d & \quad 4(d+1) \\
\end{align*}
\]

- the only $d$-type is a path of $2d + 1$ nodes
- $I_d$ and $J_d$ are $d$-equivalent

---

2-Colourability

**Theorem 11.17:** 2-Colourability is not FO-definable.

**Proof:** counterexample for quantifier rank $r$: set $d = 3^r$ (odd number)

\[
\begin{align*}
I_d & \quad J_d \\
6d & \quad 3d \\
3d & \quad 4d+2 \\
\end{align*}
\]

- the only $d$-type is a path of $2d + 1$ nodes
- $I_d$ and $J_d$ are $d$-equivalent

---

Acyclicity

**Theorem 11.18:** Acyclicity is not FO-definable.

**Proof:** counterexample for quantifier rank $r$: set $d = 3^r$

\[
\begin{align*}
I_d & \quad J_d \\
6d & \quad 3d \\
3d & \quad 4d+2 \\
\end{align*}
\]

- $d$-types are paths of $\leq 2d + 1$ nodes
- $I_d$ and $J_d$ are $d$-equivalent
Summary: Limits of FO-Queries

FO queries (and hence Relational Calculus) cannot express properties that require a "global" view:

- properties where one needs to follow paths
- properties where one needs to count elements

Remember Lecture 1?

"Stops at distance 2 from Helmholtzstr."

\[ R_2 = \delta_{\text{From}}(\pi_{\text{From}}(\text{Connect} \bowtie R_1)) \]

What about all stops reachable from Helmholtzstr.? ~ not expressible in Relational Calculus

Yet, all examples we saw are in P

~ Is there another query language that could help us?

Summary and Outlook

FO-queries (and thus CQs) cannot express even all tractable query mappings ~ FO-definability

Showing that a query is not FO-definable requires some creativity
~ Ehrenfeucht-Fraïssé Games as one approach

FO-queries can only express "local" properties

Possible proof techniques:

- Ehrenfeucht-Fraïssé Games
- Locality Theorems
- For more approaches see Chapter 17 of [Abiteboul, Hull, Vianu 1994]

Open questions:

- If FO cannot express all tractable queries, what can?