Logical foundations of databases

Diego Figueira

Gabriele Puppis

CNRS  LaBRI

day 5
Recap

- Acyclic Conjunctive Queries
- Join Trees
- Evaluation of ACQ (LOGCFL-complete)
- Ears, GYO algorithm for testing acyclicity
- Tree decomposition, tree-width of CQ
- Evaluation of bounded tree-width CQs (LOGCFL-complete)
- Bounded variable fragment of FO, evaluation in PTIME
Ehrenfeucht-Fraïssé games

They play for \( n \) rounds on the board \((S_1, S_2)\).

At each round \( i \):
- **Spoiler** chooses a node \( x_i \) from \( S_1 \) (resp. \( y_i \) from \( S_2 \))
- **Duplicator** answers with a node \( y_i \) from \( S_2 \) (resp. \( x_i \) from \( S_1 \))
  trying to maintain an isomorphism between \( S_1 \mid \{x_i\}_i \) and \( S_2 \mid \{y_i\}_i \)
Ehrenfeucht-Fraïssé games

On non-isomorphic finite structures, Spoiler wins eventually... Why?

...and he often wins very quickly:

But there are non-isomorphic infinite structures where Duplicator can survive for arbitrarily many rounds (not necessarily forever!)

Given \( n \), at each round \( i = 1, \ldots, n \), pairs of marked nodes in \( S_1 \) and \( S_2 \) must be either at equal distance or at distance \( \geq 2^{n-i} \).
Ehrenfeucht-Fraïssé games

**Theorem.** $S_1$ and $S_2$ are $n$-equivalent \([\text{Fraïssé '50, Ehrenfeucht '60}]\)

iff Duplicator has a strategy to survive $n$ rounds in the EF game on $S_1$ and $S_2$.

Proof ideas for the if-direction (from Duplicator’s winning strategy to $n$-equivalence)

Consider $\phi$ with quantifier rank $n$. Suppose $S_1 \models \phi$ and Duplicator survives $n$ rounds on $S_1, S_2$. We need to prove that $S_2 \models \phi$.

💡 A new game to evaluate formulas....
The semantics game

Assume w.l.o.g. that $\phi$ is in **negation normal form**.

Push negations inside:

$$
\neg \forall \phi \iff \exists \neg \phi \\
\neg \exists \phi \iff \forall \neg \phi \\
\neg (\phi \land \psi) \iff \neg \phi \lor \neg \psi \\
... 
$$

Whether $S \models \phi$ can be decided by a **new game** between two players, True and False:

- $\phi = E(x,y)$ $\rightarrow$ True wins if nodes marked $x$ and $y$ are connected by an edge, otherwise he loses
- $\phi = \exists x \ \phi'(x)$ $\rightarrow$ True moves by marking a node $x$ in $S$, the game continues with $\phi'$
- $\phi = \forall y \ \phi'(y)$ $\rightarrow$ False moves by marking a node $y$ in $S$, the game continues with $\phi'$
- $\phi = \phi_1 \lor \phi_2$ $\rightarrow$ True moves by choosing $\phi_1$ or $\phi_2$, the game continues with what he chose
- $\phi = \phi_1 \land \phi_2$ $\rightarrow$ False moves by choosing $\phi_1$ or $\phi_2$, the game continues with what he chose
- ...

**Lemma.** $S \models \phi$ iff True wins the semantics game.
**Theorem.** $S_1$ and $S_2$ are $n$-equivalent if and only if Duplicator has a strategy to survive $n$ rounds in the EF game on $S_1$ and $S_2$.

[Fraïssé '50, Ehrenfeucht '60]

Proof ideas for the if-direction (from Duplicator’s winning strategy to $n$-equivalence)

Consider $\phi$ with quantifier rank $n$. Suppose $S_1 \models \phi$ and Duplicator survives $n$ rounds on $S_1, S_2$. We need to prove that $S_2 \not\models \phi$.

True wins the game on $S_1$ and $S_2$.

Turn winning strategy for True in $S_1$ into winning strategy for True in $S_2$....
Proof ideas for the if-direction (from Duplicator’s winning strategy to $n$-equivalence)

Consider $\phi$ with quantifier rank $n$.

Suppose $S_1 \models \phi$ and Duplicator survives $n$ rounds on $S_1, S_2$.

We need to prove that $S_2 \not\models \phi$.

**Theorem.** $S_1$ and $S_2$ are $n$-equivalent

iff Duplicator has a strategy to survive $n$ rounds in the EF game on $S_1$ and $S_2$.

[Fraïssé '50, Ehrenfeucht '60]
Definability in FO

**Theorem.** $S_1$ and $S_2$ are *n*-equivalent if
d
Duplicator has a strategy to survive $n$ rounds in the EF game on $S_1$ and $S_2$.

**Corollary.** A property $P$ is *not definable in FO* if

\[ \forall n \exists S_1 \in P \exists S_2 \not\in P \text{ Duplicator can survive } n \text{ rounds on } S_1 \text{ and } S_2. \]

Example: $P = \{ \text{connected graphs} \}$. Given $n$, take $S_1 \in P$ large enough and $S_2 = S_1 \cup S_1 \not\in P$.
Ehrenfeucht-Fraïssé games

Several properties can be proved to be not FO-definable:

- connectivity  
  (previous slide)

- even / odd size
  
  Your turn now! ...given $n$, take $S_1 = \text{large even structure}$
  $S_2 = \text{large odd structure}$...

- 2-colorability
  
  Given $n$, take $S_1 = \text{large even cycle}$ $S_2 = \text{large odd cycle}$

- finiteness

- acyclicity

...
A different perspective: a coarser view on expressiveness...

What percentage of graphs verify a given FO sentence?
\( \mu_n(P) = \"probability that property P holds in a random graph with n nodes\" \)

\[ C_n = \{ \text{graphs with } n \text{ nodes} \} \]

\[ \mu_n(P) = \frac{\left| \{ G \in C_n \mid G \models P \} \right|}{\left| C_n \right|} = \frac{2^{n^2}}{2^{n^2}} \]

E.g. for \( P = \"the graph is complete\" \)

\[ \mu_3(P) = \frac{1}{|C_3|} = \frac{1}{2^{3^2}} \]

\[ \mu_\infty(P) = \lim_{n \to \infty} \mu_n(P) \]
Theorem. \([\text{Glebskii et al. '69, Fagin '76}]\)

For every \(FO\) sentence \(\phi\), \(\mu_{\infty}(\phi)\) is either 0 or 1.

Examples:

- \(\phi = \text{“there is a triangle”}\)
  \[\mu_3(\phi) = \frac{1}{|C_3|}, \quad \mu_{3n}(\phi) \geq 1 - \left(1 - \frac{1}{|C_3|}\right)^n \rightarrow 1\]

- \(\phi_H = \text{“there is an occurrence of } H \text{ as induced sub-graph”}\)
  \[\mu_{\infty}(\phi_H) = 1\]

- \(\phi = \text{“there no 5-clique”}\)
  \[\mu_{\infty}(\phi) = 0\]

- \(\phi = \text{“even number of edges”}\)
  \[\mu_{\infty}(\phi) = \frac{1}{2}\]

- \(\phi = \text{“even number of nodes”}\)
  \[\mu_{\infty}(\phi) \text{ not even defined}\]

- \(\phi = \text{“more edges than nodes”}\)
  \[\mu_{\infty}(\phi) = 1\] (yet not FO-definable!)
0-1 Law

For every FO sentence $\phi$, $\mu_\infty(\phi)$ is either 0 or 1.

Let $k = \text{quantifier rank of } \phi$

$$\delta_k = \forall x_1, \ldots, x_k \forall y_1, \ldots, y_k \exists z \ \land_{i,j} x_i \neq y_j \land E(x_i, z) \land \neg E(y_j, z)$$

(Extension Formula/Axiom)

Fact 1: If $G \models \delta_k \land H \models \delta_k$ then Duplicator survives $k$ rounds on $G, H$

Fact 2: $\mu_\infty(\delta_k) = 1$

($\delta_k$ is almost surely true)

2 cases

a) There is $G$ $G \models \delta_k \land \phi$ $\Rightarrow$ (by Fact 1) $\forall H :$ If $H \models \delta_k$ then $H \models \phi$

Thus, $\mu_\infty(\delta_k) \leq \mu_\infty(\phi)$

$\Rightarrow$ (by Fact 2) $\mu_\infty(\delta_k) = 1$, hence $\mu_\infty(\phi) = 1$

b) There is no $G \models \delta_k \land \phi$ $\Rightarrow$ (by Fact 2) there is $G \models \delta_k$, $G \models \delta_k \land \neg \phi$ $\Rightarrow$ (by case a) $\mu_\infty(\neg \phi) = 1$
For every FO sentence \( \phi \), \( \mu_\infty(\phi) \) is either 0 or 1, and this depends on whether \( \text{RADO} \models \phi \).

RADO =

- Each pair of nodes \( i, j \) is connected if the \( i \)-th bit of \( j \) is 1.
- Each pair of nodes \( i, j \) is connected with probability \( 1/2 \).
- The unique graph that satisfies \( \delta_k \) for all \( k \).
**Theorem.** The problem of deciding whether an FO sentence is \textit{almost surely true} ($\mu_\infty = 1$) is PSPACE-complete. [Grandjean ’83]

**Query evaluation on large databases:**
Don’t bother evaluating an FO query, it’s either \textit{almost surely true} or \textit{almost surely false}!
Does the 0-1 Law apply to real-life databases?
Not quite: database *constraints* easily spoil Extension Axiom.

Consider:

- functional constraint \( \forall x, x', y, y' \left( E(x,y) \land E(x',y') \Rightarrow y = y' \right) \land \left( E(x,y) \land E(x',y) \Rightarrow x = x' \right) \) (\( E \) is a permutation)
- FO query \( \phi = \neg \exists x \ E(x,x) \)

Probability that a permutation \( E \) satisfies \( \phi = \frac{!n}{n!} \to e^{-1} = 0.3679... \)

0-1 Law only applies to *unconstrained* databases...
Another technique: Locality

Idea: First order logic can only express “local” properties

Local = properties of nodes which are close to one another
Definition. The **Gaifman graph** of a structure $S = (V, R_1, \ldots, R_m)$ is the **undirected** graph $G_S = (V, E)$ where $E = \{(u, v) \mid \exists \ldots, u, \ldots, v, \ldots \in R_i \text{ for some } i\}$

<table>
<thead>
<tr>
<th>Agent</th>
<th>Name</th>
<th>Drives</th>
<th>Car</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>007</td>
<td>James Bond</td>
<td>Aston Martin</td>
<td>UK</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>Mr Smith</td>
<td>Cadillac</td>
<td>USA</td>
<td></td>
</tr>
<tr>
<td>201</td>
<td>Mrs Smith</td>
<td>Mercedes</td>
<td>Germany</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Jason Bourne</td>
<td>BMW</td>
<td>BMW</td>
<td>Germany</td>
</tr>
</tbody>
</table>

The Gaifman graph of a graph $G$ is the underlying undirected graph.
### Hanf locality

- \( \text{dist}(u, v) = \text{distance} \) between \( u \) and \( v \) in the Gaifman graph
- \( S[u, r] = \text{sub-structure induced by} \ \{ v \mid \text{dist}(u, v) \leq r \} = \text{ball around} \ u \ \text{of radius} \ r \)

<table>
<thead>
<tr>
<th>Agent</th>
<th>Name</th>
<th>Drives</th>
</tr>
</thead>
<tbody>
<tr>
<td>007</td>
<td>James Bond</td>
<td>Aston Martin</td>
</tr>
<tr>
<td>200</td>
<td>Mr Smith</td>
<td>Cadillac</td>
</tr>
<tr>
<td>201</td>
<td>Mrs Smith</td>
<td>Mercedes</td>
</tr>
<tr>
<td>3</td>
<td>Jason Bourne</td>
<td>BMW</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Car</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aston Martin</td>
<td>UK</td>
</tr>
<tr>
<td>Cadillac</td>
<td>USA</td>
</tr>
<tr>
<td>Mercedes</td>
<td>Germany</td>
</tr>
<tr>
<td>BMW</td>
<td>Germany</td>
</tr>
</tbody>
</table>

---

![Diagram](diagram.png)
**Hanf locality**

Definition. Two structures $S_1$ and $S_2$ are **Hanf** $(r, t)$-equivalent iff for each structure $B$, the two numbers

$$\#u \text{ s.t. } S_1 [u, r] \cong B \quad \#v \text{ s.t. } S_2 [v, r] \cong B$$

are *either the same* or *both $\geq t$*.

Example. $S_1, S_2$ are Hanf $(1, 1)$-equivalent iff they have the *same balls* of radius 1.
Definition. Two structures $S_1$ and $S_2$ are **Hanf ($r, t$) - equivalent** iff for each structure $B$, the two numbers

$$
\# u \text{ s.t. } S_1[u, r] \cong B \quad \# v \text{ s.t. } S_2[v, r] \cong B
$$

are *either the same* or *both* $\geq t$.

Example. $K_n, K_{n+1}$ are **not** Hanf ($1, 1$) - equivalent
Hanf locality

**Theorem.** If $S_1, S_2$ are Hanf $(r, t)$-equivalent, with $r = 3^n$ and $t = n$ then $S_1, S_2$ are $n$-equivalent (they satisfy the same sentences with quantifier rank $n$)

[Hanf '60]

**Exercise:** prove that *acyclicity* is not FO-definable (on finite structures)
**Theorem.** $S_1, S_2$ are $n$-equivalent (they satisfy the same sentences with quantifier rank $n$) whenever $S_1, S_2$ are Hanf $(r, t)$-equivalent, with $r = 3^n$ and $t = n$. [Hanf '60]

**Exercise:** prove that testing whether a binary tree is *complete* is not FO-definable
**Theorem.** $S_1, S_2$ are $n$-equivalent (they satisfy the same sentences with quantifier rank $n$) whenever $S_1, S_2$ are Hanf $(r, t)$-equivalent, with $r = 3^n$ and $t = n$.

[Hanf '60]

Why so BIG?

Remember $\phi_k(x,y) = "\text{there is a path of length } 2^k \text{ from } x \text{ to } y"$

$\phi_0(x, y) = E(x, y)$, and

$\phi_k(x,y) = \exists z \left( \phi_{k-1}(x, z) \land \phi_{k-1}(z, y) \right)$

$qr(\phi_k) = k$

Not $(n+2)$-equivalent yet they have the same $2^{n-1}$ balls.
Gaifman locality

What about queries?

Eg: Is reachability expressible in FO?

What about equivalence on the same structure?
When are two points indistinguishable?
Gaifman locality

\[ S[(a_1, ..., a_n), r] = \text{induced substructure of } S \]

of elements at distance \( \leq r \) of some \( a_i \) in the Gaifman graph.
Gaifman locality

For any $\phi \in \text{FO}$ of quantifier rank $k$ and structure $S$,

$$S \left[ (a_1, \ldots, a_n), r \right] \cong S \left[ (b_1, \ldots, b_n), r \right] \quad \text{for } r = 3^{k+1}$$

implies

$$(a_1, \ldots, a_n) \in \phi(S) \iff (b_1, \ldots, b_n) \in \phi(S)$$

Idea: If the neighbourhoods of two tuples are the same, the formula cannot distinguish them.
Gaifman locality vs Hanf locality

Difference between Hanf- and Gaifman-locality:

Hanf-locality relates **two different structures**,

$S_1$ and $S_2$ have the same # of balls of radius $3^k$, **up to threshold $k$**

\[ \downarrow \]

They verify the same sentences of $qr \leq k$

Gaifman-locality talks about definability in **one structure**

Inside $S$,

$3^{k+1}$-balls of $(a_1,\ldots,a_n) = 3^{k+1}$-balls of $(b_1,\ldots,b_n)$

\[ \downarrow \]

$(a_1,\ldots,a_n)$ indistinguishable from $(b_1,\ldots,b_n)$ through **formulas** of $qr \leq k$
A query $Q(x_1, ..., x_n)$ is not FO-definable if:

- for every $k$ there is a structure $S_k$ and $(a_1, ..., a_n), (b_1, ..., b_n)$ such that
  - $S_k [(a_1, ..., a_n), 3^{k+1}] \cong S_k [(b_1, ..., b_n), 3^{k+1}]$
  - $(a_1, ..., a_n) \in Q(S_k)$, $(b_1, ..., b_n) \not\in Q(S_k)$

Proof: If $Q$ were expressible with a formula of quantifier rank $k$, then $(a_1, ..., a_n) \in Q(S_k)$ iff $(b_1, ..., b_n) \in Q(S_k)$. Absurd!
Gaifman locality

Reachability is not FO definable.

For every $k$, we build $S_k$:

And $S_k [(a_1, a_2), 3^{k+1}] \equiv S_k [(b_1, b_2), 3^{k+1}]$

However,

- $b_2$ is reachable from $b_1$,
- $a_2$ is not reachable from $a_1$.

Your turn! $Q(x) =$ “$x$ is a vertex separator”
Gaifman Theorem

Basic local sentence:

\[ \exists x_1, \ldots, x_n \]

\[ r \]

\[ x_1 \]

\[ x_2 \]

\[ \cdots \]

\[ x_n \]

disjoint \( r \)-balls around \( x_1, \ldots, x_n \)

\[ \land \psi_1(x_1) \land \cdots \land \psi_n(x_n) \]

\( r \)-local formulas

Inside \( \psi_i(x_i) \) we interpret \( \exists y. \phi \) as \( \exists y. d(x_i, y) \leq r \land \phi \)

Gaifman Theorem: Every FO sentence is equivalent to a boolean combination of basic local sentences.
Recap

**EF games**

FO sentences with quantifier rank n = winning strategies for Spoiler in the n-round EF game

**0-1 Law**

FO sentences are almost always true or almost always false

**Hanf locality**

FO sentences with quantifier rank n = counting $3^n$ sized balls up to n

**Gaifman locality**

Queries of quantifier rank n output tuples closed under $3^{n+1}$ balls.

**Gaifman Theorem**

An FO sentence can only say

“there are some points at distance $\geq 2r$ whose r-balls are isomorphic to certain structures”

or a boolean combination of that.
Some more cool stuff...

Descriptive complexity

What properties can be checked efficiently? E.g. 3COL can be tested in NP

Metatheorem

“A property can be expressed in [insert some logic here] iff it can be checked in [some complexity class here]”

⇒ “A property is FO-definable iff it can be tested in AC$^0$”

⇒ “A property is $\exists$SO-definable iff it can be tested in NP” [Fagin 73]

⇒ Open problem: which logic captures PTIME?
Some more cool stuff...

Recursion

Can we enhance query languages with recursion? E.g. express reachability properties

\[
\text{Ancestor}(X,Y) \leftarrow \text{Parent}(X,Z), \text{Ancestor}(Z,Y)
\]

\[
\text{Ancestor}(X,X) \leftarrow .
\]

\(?- \text{Ancestor}(“Louis XIV”,Y)\)

\[\Rightarrow\] Incomparable with FO (has recursion, but is monotone)

\[\Rightarrow\] Evaluation is in PTIME (for data complexity, but also for bounded arity)

Datalog (semantics based on least fixpoint)
Some more cool stuff...

**Semi-structured data**

Tree-structured or graph-structures dbs in place of relational dbs.

XML, XPath, Stream processing, ...

```xml
<catalog>
  <book id="1">
    <title>XML Developer's Guide</title>
    <author>Matthew Gambardella</author>
    <year>2000</year>
  </book>
  <book id="2">
    <title>Beginning XML</title>
    <author>David Hunter</author>
    <author>David Gibbons</author>
    <year>2007</year>
  </book>
  ...
</catalog>
```

⇝ Evaluation of XPath is in linear time (data complexity)  
[Bojanczyk, Parys 08]

⇝ Satisfiability for $\text{FO}^2[\downarrow, \neg]$ is decidable  
[Bojanczyk, Muscholl, Schwentick, Segoufin 09]
Some more cool stuff...

Incomplete information

How to correctly reason when information is hidden/missing/noisy/...?

Certain Query Answers (CQA)

\[ \phi[V] = \bigcap_{D \in [V]} \phi(D) \]

\[ \rightarrow \text{ CQA computable in PTIME w.r.t. view size.} \quad [\text{Abiteboul, Kanellakis, Grahne 91}] \]
Bibliography

  (available at http://webdam.inria.fr/Alice/)


  (available at www.mathematik.tu-darmstadt.de/~otto/LEHRE/FMT0809.ps)

  (available at www.math.helsinki.fi/logic/people/jouko.vaaninen/shortcourse.pdf)