Expressive Completeness of Two-Variable First-Order Logic with Counting for First-Order Logic Queries on Rooted Unranked Trees

Jelle Hellings — McMaster University
Marc Gyssens — Hasselt University
Jan Van den Bussche — Hasselt University
Dirk Van Gucht — Indiana University

Main Result

Let φ be an unary *first-order* query.
There exists an FO^2+C query ψ that is equivalent to φ on *node-labeled, unranked, unordered trees*.
(Here, we ignore node labels.)

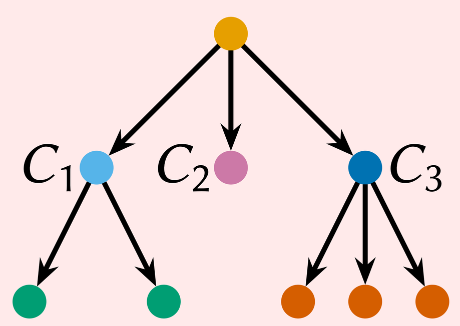
FO^2+C Queries on Trees

Let φ be a unary first-order query and $\mathcal{T}$ a tree.

1. There exists an unary FO^2+C query $\mathrm{tq}_{\mathcal{T}}$ with
$$[\mathrm{tq}_{\mathcal{T}}]_{\mathcal{T}'} \neq \emptyset$$
if and only if trees $\mathcal{T}$ and $\mathcal{T}'$ are isomorphic.
2. There exists an unary FO^2+C query $\mathrm{tn}_{\mathcal{T}}$ with
$$[\mathrm{tn}_{\mathcal{T}}]_{\mathcal{T}} = [\varphi]_{\mathcal{T}}.$$
3. Let $\mathbb{T}$ be the set of all trees. The query φ is equivalent to FO^2+C query
$$Q_\varphi := \bigvee_{\mathcal{T}' \in \mathbb{T}} \Big((\exists v\ (\mathrm{tq}_{\mathcal{T}'})) \wedge \mathrm{tn}_{\mathcal{T}'} \Big).$$

Main challenge Restrict $\mathbb{T}$ to a *finite set*.

Example



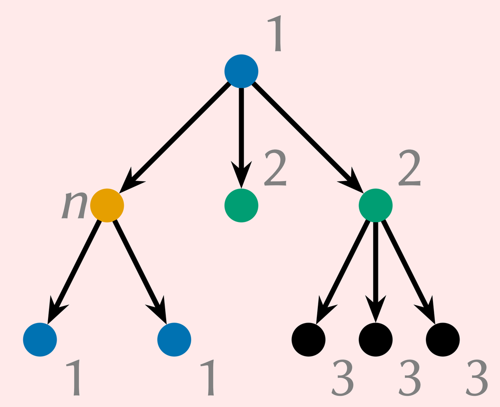
$$(\exists^{=1} v\ (\mathrm{root}(v) \wedge (\exists^{=3} w\ \mathrm{edge}(v, w)) \wedge C_1 \wedge C_2 \wedge C_3),$$
$$C_1 := \exists^{=1} w\ (\mathrm{edge}(v, w) \wedge (\exists^{=2} v\ \mathrm{edge}(w, v)) \wedge (\exists^{=2} v\ \mathrm{edge}(w, v) \wedge \mathrm{leaf}(v)));$$
$$C_2 := \exists^{=1} w\ (\mathrm{edge}(v, w) \wedge \mathrm{leaf}(w));$$
$$C_3 := \exists^{=1} w\ (\mathrm{edge}(v, w) \wedge (\exists^{=3} v\ \mathrm{edge}(w, v)) \wedge (\exists^{=3} v\ \mathrm{edge}(w, v) \wedge \mathrm{leaf}(v))).$$

Using Hanf Locality

Let $\mathcal{T} = (\mathcal{N}, \mathcal{E})$ be a tree and let $n \in \mathcal{N}$.

- The *d-neighborhood* around n is the set of nodes (subtree) reachable from n via a path of at-most d edges.
- Two trees are *(d, m)-equivalent* if they have the *same amount* (up-till-m) of each d-neighborhood.

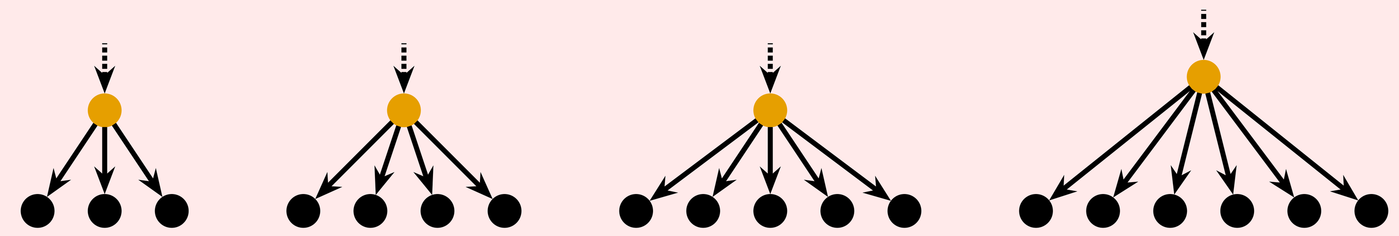
Example



Theorem (Fagin et al.)

1. *If every node has at-most f children, then there is a finite number of distinct d-neighborhoods (up-to-isomorphisms).*
2. *If every node has at-most f children, then there exists d, m that only depend on r, f such that if two trees are (d, m)-equivalent, then they are indistinguishable by r-round EF-games.*

Limitations on *unranked* trees



All four nodes have distinct d-neighborhoods, $d \geq 1$.

Our main technical contribution

A first-order locality notion that takes into account *branching* and is expressible in FO^2+C.

Bounded Equivalence on *Nodes* and *Trees*

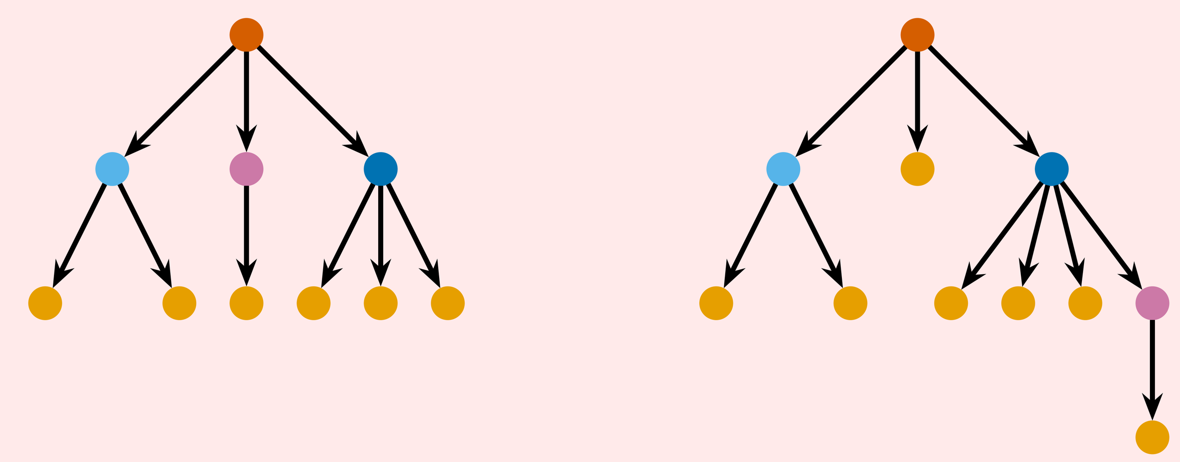
Let $\mathcal{T}_1 = (\mathcal{N}_1, \mathcal{E}_1)$, $\mathcal{T}_2 = (\mathcal{N}_2, \mathcal{E}_2)$ be two trees, and let $n_1 \in \mathcal{N}_1$, $n_2 \in \mathcal{N}_2$.

Nodes n_1 and n_2 are *downward (b, d)-bounded equivalent* ($n_1 \approx_{\downarrow b,d} n_2$) if either $d = 0$ or their children can be grouped into equivalence classes based on $\approx_{\downarrow b,d-1}$, and these classes have *the same size* (up-till-b).

Nodes n_1 and n_2 are *(b, d)-bounded equivalent* if $n_1 \approx_{\downarrow b,d} n_2$ and their parents (if any) are $(b, d-1)$-bounded equivalent.

Example

(3, 1)-bounded equivalence classes



Trees $\mathcal{T}_1$ and $\mathcal{T}_2$ are *(b, d, k)-bounded equivalent* if "they are similar" with respect to sets-of-at-most-k nodes that are (b, d)-bounded equivalent.

Theorem

1. *The above notions are FO^2+C expressible.*
2. *There exists a finite number of distinct (b, d)-bounded equivalence classes.*
3. *Let $r \geq 0$, and $d = 7^r - 1$, $b = r + 2$, $k = 4d + 4$. If $\mathcal{T}_1 \approx_{b,d,k} \mathcal{T}_2$ and $n_1 \approx_{b,d} n_2$, then n_1 and n_2 are indistinguishable by r-round EF-games.*