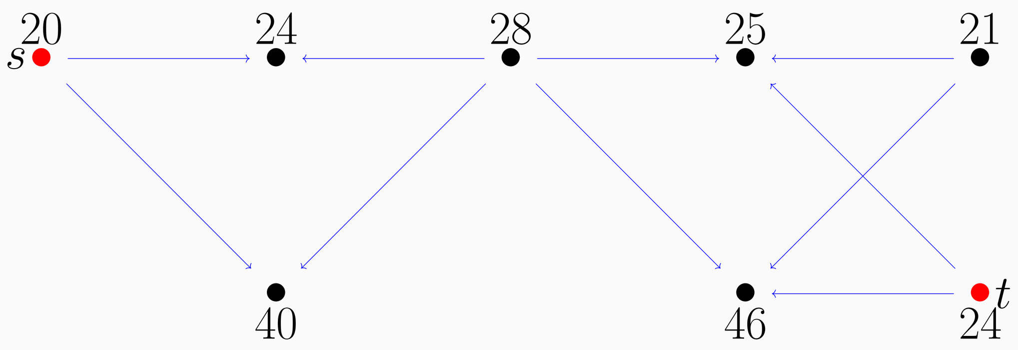
# Extending Pattern Matching Queries in Property Graphs with Interpreted Predicates

Leonid Libkin[2,3,5], Cristina Sirangelo[1,2,4], and Deniz Yilmaz[1,2,4]

[1] CNRS [2]IRIF [3]RelationalAI [4]Université Paris Cité [5]University of Edinburgh

## Example 1

Social network: nodes are users carrying age information and $\rightarrow$ is "follow".




$u, v$ are *potential friends* if

- there is $w$ such that $u \rightarrow w \leftarrow v$ and
- both have less than five years age difference with $w$.

**Query:** *Are $s$ and $t$ connected with a chain of potential friends?*

## Data-Path Queries

"Data-Path = Navigational + Relational"

**Candidate:** Regular Path Queries and first order logic over the underlying data structure.

**Example 1:** a typical data-path query using no unrestricted quantifiers.

**Example 2:** a typical data-path query using unrestricted quantifiers.

## core-Data-Path Logic

Key construct of cDPL is given by:

$$e_{x,y}[\varphi(x,y)](s,t)$$

where

- $e$ is a regular expression over $\Sigma$;
- $\varphi(x,y)$ is a (two-sorted) first-order formula over $\{M, G\}$;
- $s, t$ are the only free variables of the cDPL-formulas.

**Semantics**: Let $\mathcal{G} = \{M, G\}$.
$\mathcal{G} \models e_{x,y}[\varphi(x,y)](s,t)$ iff

1. there is a path $\pi$ in $G$ with the label $\lambda(\pi) \in e$ and
2. for every edge $(v_j, v_{j+1})$ in $\pi$ we have $\mathcal{G} \models_{[v_j/x,\ v_{j+1}/y]} \varphi(x,y)$.

**Example 1** can be expressed in cDPL by

$$(\rightarrow\leftarrow)^*_{x,y}[(\mathsf{age}(y) < \mathsf{age}(x)+5) \vee (\mathsf{age}(x) < \mathsf{age}(y)+5)](s,t).$$

⚠ cDPL can not express Example 2 since it requires quantification "outside" the cDPL-expressions. So we shall extend cDPL by closing it under first-order logic. (go to DPL)

## Collapse Results

1. $\mathsf{DPL}_{\mathsf{act}}$ : the fragment of DPL which uses only restricted quantifiers.
2. $M$ admits *restricted quantifier collapse* for DPL if every DPL query is equivalent to a $\mathsf{DPL}_{\mathsf{act}}$ query.

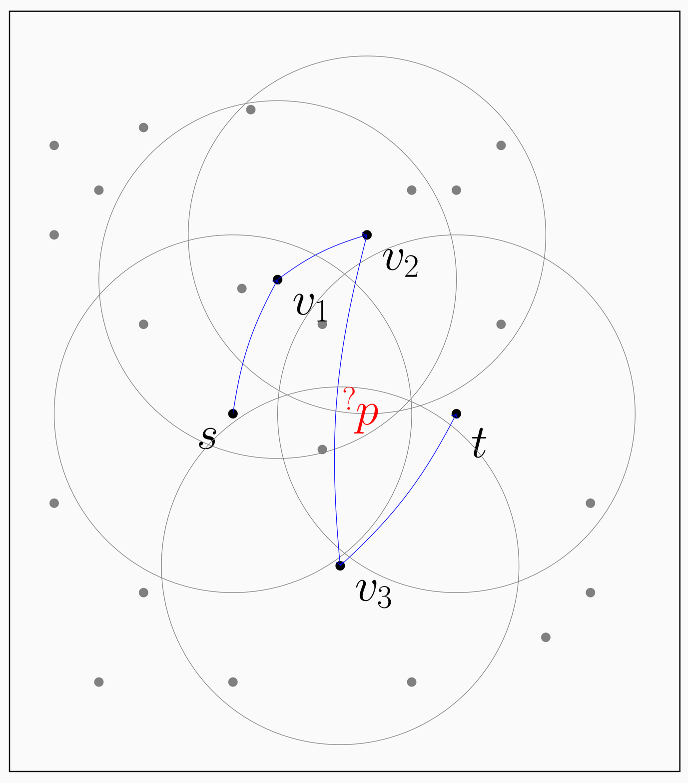
**Theorem.** *Every "good" $M$ admits restricted quantifier collapse for* DPL.

**Examples** of
*Good structures:* $(\mathbb{Q},<)$, $(\mathbb{R},<)$, $(\mathbb{Q},+,<)$, $(\mathbb{R},+,<)$, $(\mathbb{R},+,\times,<)$, …
*Bad structures:* $(\mathbb{N},+,\times)$, Random Graph, $(\mathbb{N},+,2^n)$, $(\mathbb{Q},+,\times,<)$ …

⚠ "Good" means $M$ is *o-minimal* and admits *quantifier elimination.*

## Example 2

Graph on the plane: nodes are points on $\mathbb{R}^2$ and $\Sigma$-labeled edges.




**Query:** *Is there a path $\pi$ from $s$ to $t$ such that every node in $\pi$ is at distance less than 1 from a common point $p \in \mathbb{R}^2$?*

⚠ $p$ does not need to be a node in the graph.

## Structures

An *$n$-dimensional embedded data graph* consists of:

1. an underlying structure $M$, for data values and operations on them, and
2. a finite graph $G$ whose nodes are $n$-tuples over $M$ and whose edges are labeled by elements of a finite alphabet $\Sigma$.

**Examples** of underlying structures: $(\mathbb{Q},<)$, $(\mathbb{R},+,<)$, $(\mathbb{R},+,\times,<)$, $(\mathbb{Z},+,<)$ …

⚠ This framework is an instance of the "Embedded Finite Structures" introduced in Benedikt and Libkin, 1996.

## Data Path Logic

The syntax of *Data-Path Logic*, denoted DPL is given by:

$$e_{x,y}[\varphi(x,y,\bar{z})](s,t) \mid \neg\Phi \mid \Phi \vee \Psi \mid \Phi \wedge \Psi \mid \exists z\, \Phi \mid \forall z\, \Phi$$

where the atoms $e_{x,y}[\varphi(x,y,\bar{z})](s,t)$ as cDPL, but allow additional free variables.

⚠ These free variables may be of sort $G$ or $M$. Thus, quantification may range either

- over nodes of $G$ (called active domain quantification) or
- over elements of $M$ (called unrestricted quantification).

**Example 2** can be expressed in DPL by

$$\exists p_1 p_2\, \Sigma^*_{x,y}[(x_1-p_1)^2 + (x_2-p_2)^2 < 1 \wedge (y_1-p_1)^2 + (y_2-p_2)^2 < 1](s,t).$$

**Q:** Can we evaluate DPL?
**A:** It depends on the model theory.

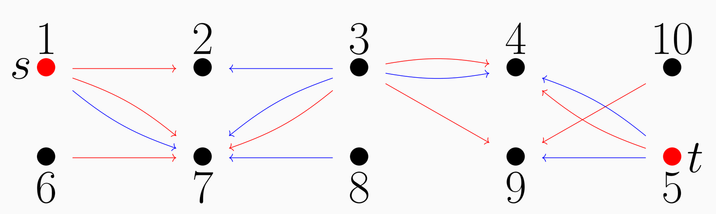
## Complexity Results

**Proposition.** *If $M$ is "good", then:*

1. *Data complexity of* DPL *is in* **NL**.
2. *Data complexity with restrictors on the shapes of paths **trail** or **acyclic** is in* **NP**.

## What's next?

Observe that DPL has very limited nesting.
**Example 3**




**Query:** *Is there a path $\pi$ from $s$ to $t$ such that*

1. *$\pi$ follows the pattern $x \rightarrow z \leftarrow y$ and*
2. *$z = \frac{x+y}{2}$?*

⚠ DPL can not express this query requiring nesting.

### Regular Expressions with Conditions

$$a \mid e.e' \mid e \cup e' \mid e^- \mid e^* \mid e[\varphi]$$

where $a \in \Sigma$, $e, e' \in \mathsf{REC}$ and $\varphi \in \mathcal{FO}(M,\Sigma)$.

**Example 3** can be expressed in REC:

$$\Big((\rightarrow\leftarrow)[\exists z(x \rightarrow z) \wedge (z \leftarrow y) \wedge (z + z = x + y)]\Big)^*(s,t)$$

**Observations:**

- cDPL is a sublogic of REC.
- DPL is a sublogic of $\mathcal{FO}$(REC).
- $\mathcal{FO}$(REC) has the same collapse results and data complexity as DPL.

⚠ cDPL is "flat" REC.

$\mathcal{G} \models e(s,t)$ iff there exists $\pi^t_s \Vdash e$

where we define $\Vdash$ recursively as:

1. $\pi^t_s \Vdash a$ iff $(s,t) \in a^{\mathcal{G}}$.
2. $\pi^t_s \Vdash e.e'$ iff there exists $\pi_1, \pi_2$ such that $\pi^t_s = \pi_1\pi_2$ and $\pi_1 \Vdash e$ and $\pi_2 \Vdash e'$.
3. $\pi^t_s \Vdash e \cup e'$ iff $\pi^t_s \Vdash e$ or $\pi^t_s \Vdash e'$.
4. $\pi^t_s \Vdash e^-$ iff $\pi^s_t \Vdash e$.
5. $\pi^t_s \Vdash e^*$ iff for some $n$, there exists $\pi_1, \dots, \pi_n$ such that $\pi^t_s = \pi_1 \dots \pi_n$ and $\pi_i \Vdash e$ for all $1 \le i \le n$.
6. $\pi^t_s \Vdash e[\varphi]$ iff $\pi^t_s \Vdash e$ and $\mathcal{G} \models \varphi(s,t)$.

**Q.** Can $\mathcal{FO}$(REC) express the path patterns of GQL and serve to enrich it with additional data types?