Parsing Techniques for Contex-Free Path Querying

Semyon Grigorev

JetBrains Research, Programming Languages and Tools Lab
Saint Petersburg University

April 05, 2019

Programming Languages and Tools Lab

- https://research.jetbrains.org/groups/plt_lab

Formal languages for data analysis

- Semyon Grigorev

Topics of interest

- Formal language theory
- Parsing algorithms

Formal language constrained path querying

- Finite directed edge-laballed graph $\mathcal{G} = (V, E, L)$
- The path is a world over L:
  $\omega(p) = \omega(v_0 \xrightarrow{l_0} v_1 \xrightarrow{l_1} \ldots \xrightarrow{l_{n-1}} v_n) = l_0 \cdot l_1 \cdot \ldots \cdot l_{n-1}$
- The language $\mathcal{L}$ (over L)

Formal language constrained path querying

- Finite directed edge-laballed graph $\mathcal{G} = (V, E, L)$
- The path is a world over L:
  $\omega(p) = \omega(v_0 \xrightarrow{l_0} v_1 \xrightarrow{l_1} \ldots \xrightarrow{l_{n-1}} v_n) = l_0 \cdot l_1 \cdot \ldots \cdot l_{n-1}$
- The language $\mathcal{L}$ (over L)

- Reachability problem: $Q = \{(v_i, v_j) \mid \exists p = v_i \ldots v_j, \omega(p) \in \mathcal{L}\}$
- Path querying problem: $Q = \{p \mid \omega(p) \in \mathcal{L}\}$
  - Single path, all paths, shortest path ...

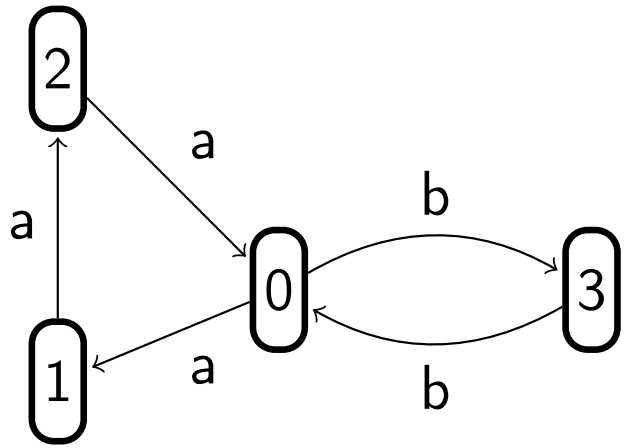
Context-Free path querying

- $\mathcal{L}$ is a context-free language
- $G_{\mathcal{L}} = (N, \Sigma, R, S)$
- Reachability problem: $Q = \{(v_i, v_j) \mid \exists p = v_i \dots v_j, S \xrightarrow[G_L]{*} \omega(p)\}$
- Path querying problem: $Q = \{p \mid \omega(p) \in \mathcal{L}\}$

Example of CFPQ

$$S \to a\ S\ b$$
$$S \to a\ b$$

(a) Grammar G_1 for $\{a^n b^n \mid n > 0\}$



(b) Input graph D_1

Paths:

$2 \xrightarrow{a} 0 \xrightarrow{b} 3$

$1 \xrightarrow{a} 2 \xrightarrow{a} 0 \xrightarrow{b} 3 \xrightarrow{b} 0$

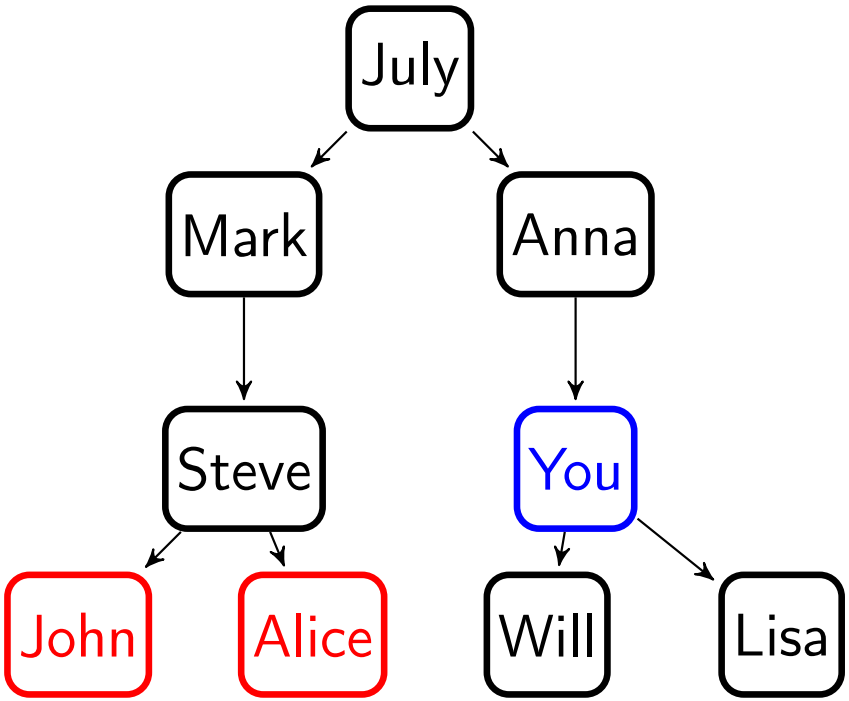
$0 \xrightarrow{a} 1 \xrightarrow{a} 2 \xrightarrow{a} 0 \xrightarrow{b} 3 \xrightarrow{b} 0 \xrightarrow{b} 3$

...

Applications

- Graph data bases querying
- Static code analysis
- Error recovery

Graph data bases querying



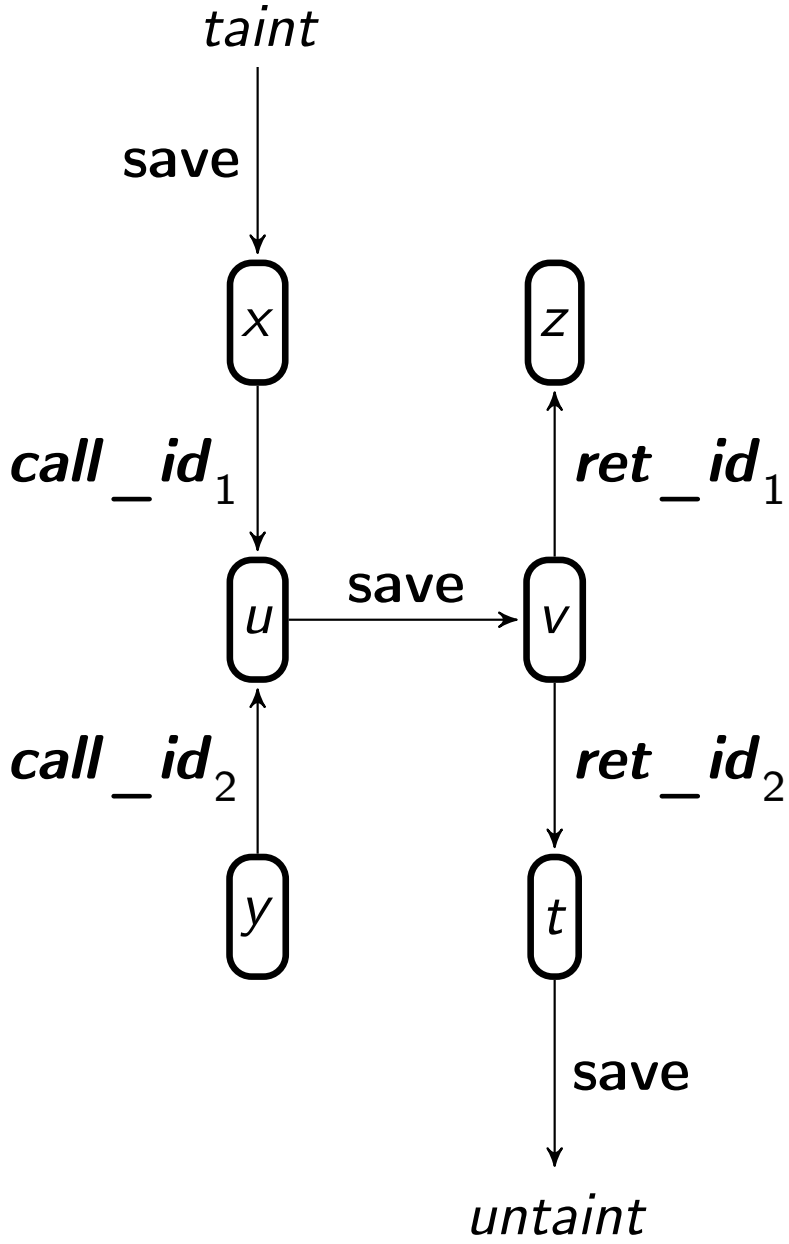
Find your cousins once removed

$$S \to H \downarrow$$
$$H \to \varepsilon \mid \uparrow H \downarrow$$

Same generation query, similarity query.

Static code analysis

```
int id(int u)
{
  v = u;
  return v;
}
int main()
{
  //taint
  int x;
  int z, y;
  //untaint
  int t;
  z = id(x);
  t = id(y);
}
```



Bioinformatics. Sequences analysis

- High performance
- New classes of grammars
- !!!!

Contact Information

- Semyon Grigorev:
  - s.v.grigoriev@spbu.ru
  - Semen.Grigorev@jetbrains.com
- Polina Lunina:
  - lunina_polina@mail.ru

- Trained models: https://github.com/YaccConstructor/YC.Bio

Thanks!