A Compendium of Regular Expr. Shapes in SPARQL Queries

Janik Hammerer **Wim Martens**

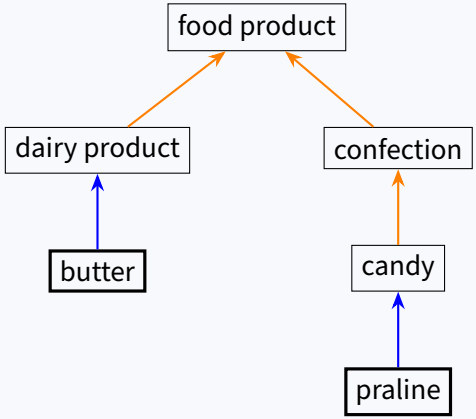
Data-Intensive Computing (AI VII)
University of Bayreuth

AlMoTh 2026

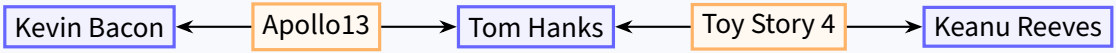
Regular Path Queries in SPARQL

```
SELECT ?var
WHERE {
 ?var wdt:instance_of /
      wdt:subclass_of +
      wd:food_product.
}
```

/	concatenation
|	disjunction
*	zero or more
+	one or more
?	zero or one
^	inverse (object to subject)



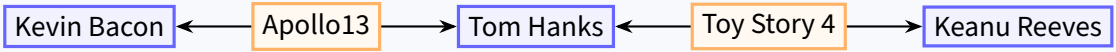
UNIVERSITÄT BAYREUTH




```
SELECT ?actor
WHERE {
   wd:Kevin_Bacon (^wdt:cast_member/wdt:cast_member)* ?actor.
}
```

?actor
Kevin Bacon
Tom Hanks
Keanu Reeves
⋮

UNIVERSITÄT BAYREUTH




```
SELECT ?actor
WHERE {
   wd:Kevin_Bacon (^wdt:cast_member/wdt:cast_member)* ?actor.
}
```

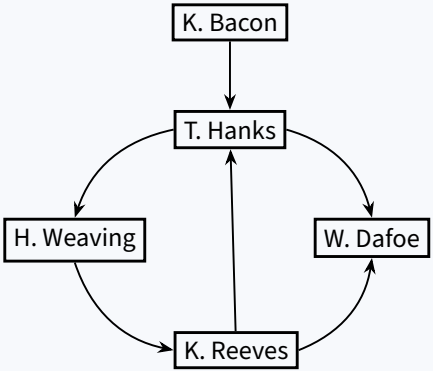
?actor	?bacon_num	?path
Kevin Bacon	?	?
Tom Hanks	?	?
Keanu Reeves	?	?
⋮	⋮	⋮

UNIVERSITÄT BAYREUTH

Path Modes (Cypher, GQL, SQL/PGQ)

- shortest path
- simple path (no repeated nodes)
- trail (no repeated edges)

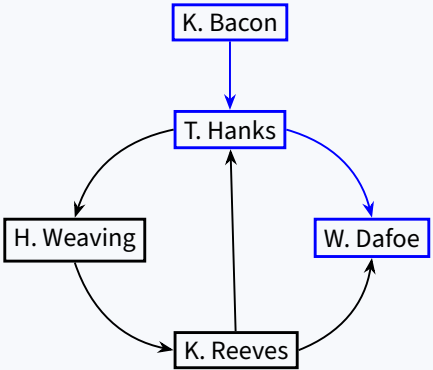
⇒ Guaranteed finite number of paths



Path Modes (Cypher, GQL, SQL/PGQ)

- shortest path
- simple path (no repeated nodes)
- trail (no repeated edges)

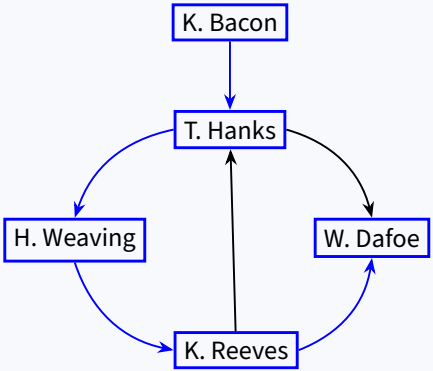
⇒ Guaranteed finite number of paths



Path Modes (Cypher, GQL, SQL/PGQ)

- shortest path
- simple path (no repeated nodes)
- trail (no repeated edges)

⇒ Guaranteed finite number of paths

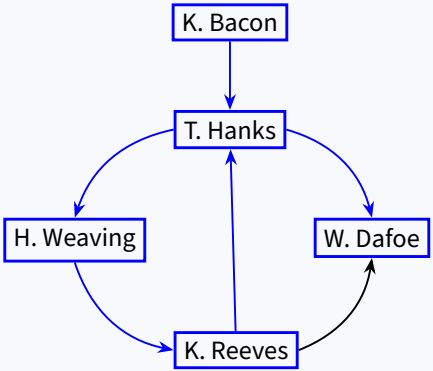


UNIVERSITÄT BAYREUTH

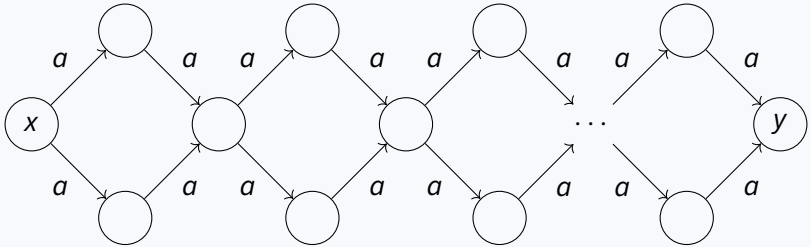
Path Modes (Cypher, GQL, SQL/PGQ)

- shortest path
- simple path (no repeated nodes)
- trail (no repeated edges)

$\Rightarrow$ Guaranteed finite number of paths



UNIVERSITÄT BAYREUTH



- $\Rightarrow$ exponentially many paths from x to y (e.g. a^*)
- $\Rightarrow$ standard suggests to store such paths separately
- $\Rightarrow$ compact representations of intermediate results needed
- $\Rightarrow$ for efficient path counting/enumeration: determinism/unambiguity

UNIVERSITÄT BAYREUTH

Research Questions

1. Syntactic Structure

What is the syntactic structure of RPQs in practice?

1. Regular expr. shapes
2. RPQ operators

UNIVERSITÄT BAYREUTH

Research Questions

1. Syntactic Structure

What is the syntactic structure of RPQs in practice?

1. Regular expr. shapes
2. RPQ operators

2. Determinism

How much nondeterminism do RPQs have in practice?

1. #deterministic RPQs
2. Size of equivalent DFAs/UFAs

3. Evaluation Efficiency

Which RPQs can be evaluated efficiently under different path modes?

1. Simple path
2. Trail

UNIVERSITÄT BAYREUTH

Data Set

Source	*#Queries*	*#Queries using RPQs*	
Wikidata (R)	572,329,407	132,303,491	**23.11 %**
Wikidata (O)	3,572,387	1,209,348	**33.85 %**
DBpedia	292,282,591	1,463,658	0.50 %
Other	27,621,127	3,286	0.01 %
Total	927,005,680	134,979,783	14.56 %

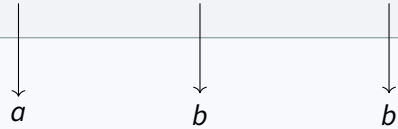
UNIVERSITÄT BAYREUTH

```
SELECT ?var1
WHERE {
   wd:Albert_Einstein (wdt:mother / wdt:child / wdt:child *) ?var1.
}
```

```
SELECT ?var1
WHERE {
   wd:Albert_Einstein (wdt:mother / wdt:child / wdt:child *) ?var1.
}
```

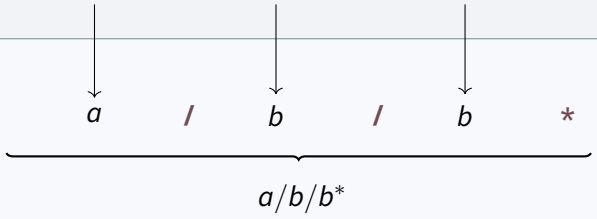
```
SELECT ?var1
WHERE {
   wd:Albert_Einstein (wdt:mother / wdt:child / wdt:child *) ?var1.
}
```

```
SELECT ?var1
WHERE {
   wd:Albert_Einstein (wdt:mother / wdt:child / wdt:child *) ?var1.
}
```

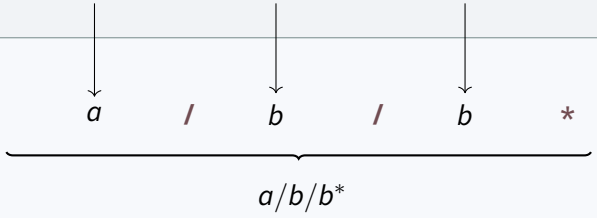


UNIVERSITÄT BAYREUTH

```
SELECT ?var1
WHERE {
   wd:Albert_Einstein (wdt:mother / wdt:child / wdt:child *) ?var1.
}
```



```
SELECT ?var1
WHERE {
   wd:Albert_Einstein (wdt:mother / wdt:child / wdt:child *) ?var1.
}
```



148.7M RPQs ⟶ **572** Regular Expr. Shapes

Regular Expression Shapes

Expr.	*#RPQ*	*Rel.*
a/b^*	1,024,661	76.28%
a^*	93,133	6.93%
a/b	65,588	4.88%
$a \mid b$	46,531	3.46%
a/b^+	11,790	0.88%
⋮	⋮	⋮
Total	1,343,268	100%

Operator Usage

Op.	*#RPQ*	*Rel.*
$*$	1.169.299	87.05 %
$/$	1.152.570	85.80 %
$\mid$	97.335	7.25 %
+	22.581	1.68 %
?	9.430	0.70 %
^	8.223	0.61 %
!	3.655	0.27 %
Total	1,343,268	100%

Dataset on Zenodo

WD_organic.csv

ID	Expression	Valid #RPQ	Valid %	Unique #RPQ	Unique %	Det?	#St	SP_tract	T_tract
1	a/(b)*	1024661	76.28%	73189	35.87%	yes	2	yes	yes
2	(a)*	93133	6.93%	43672	21.40%	yes	1	yes	yes
3	a/b	65588	4.88%	26593	13.03%	yes	3	yes	yes

- 29 CSV files
- Regular Expr. Shapes with frequency
- size of equivalent minimal DFAs
- determinism, T_{tract} and SP_{tract} classification

What you can do with this

- What proportion of real-world RPQs are FO-definable?
- What proportion of real-world RPQs are FO^2-definable?
- . . .

Determine this for 148.7M real-world RPQs

Definition (Deterministic Regular Expressions)

A regular expression is deterministic if, when reading a word from left to right, it is always clear where in the expression the current symbol can be matched without looking ahead.

Example

word $w = baa$

regex $r_1 = b(a + b)^*a$

UNIVERSITÄT BAYREUTH

Definition (Deterministic Regular Expressions)

A regular expression is deterministic if, when reading a word from left to right, it is always clear where in the expression the current symbol can be matched without looking ahead.

Example

word $w = \mathbf{b}aa$

regex $r_1 = \mathbf{b}(a + b)^*a$

UNIVERSITÄT BAYREUTH

Definition (Deterministic Regular Expressions)

A regular expression is deterministic if, when reading a word from left to right, it is always clear where in the expression the current symbol can be matched without looking ahead.

Example

word $w = b\mathbf{a}a$

regex $r_1 = b(\mathbf{a} + b)^*\mathbf{a}$

$\Rightarrow r_1$ is nondeterministic.

UNIVERSITÄT BAYREUTH

Definition (Deterministic Regular Expressions)

A regular expression is deterministic if, when reading a word from left to right, it is always clear where in the expression the current symbol can be matched without looking ahead.

Example

word $w = baa$

regex $r_1 = b(a + b)^*a$

$\Rightarrow r_1$ is nondeterministic.

regex $r_2 = bb^*a(b^*a)^*$

UNIVERSITÄT BAYREUTH

Definition (Deterministic Regular Expressions)

A regular expression is deterministic if, when reading a word from left to right, it is always clear where in the expression the current symbol can be matched without looking ahead.

Example

word $w = \mathbf{b}aa$

regex $r_1 = b(a + b)^*a$

$\Rightarrow r_1$ is nondeterministic.

regex $r_2 = \mathbf{b}b^*a(b^*a)^*$

UNIVERSITÄT BAYREUTH

Definition (Deterministic Regular Expressions)

A regular expression is deterministic if, when reading a word from left to right, it is always clear where in the expression the current symbol can be matched without looking ahead.

Example

word $w = b\mathbf{a}a$

regex $r_1 = b(a + b)^*a$

$\Rightarrow r_1$ is nondeterministic.

regex $r_2 = bb^*\mathbf{a}(b^*a)^*$

UNIVERSITÄT BAYREUTH

Definition (Deterministic Regular Expressions)

A regular expression is deterministic if, when reading a word from left to right, it is always clear where in the expression the current symbol can be matched without looking ahead.

Example

word $w = ba\mathbf{a}$

regex $r_1 = b(a + b)^*a$

$\Rightarrow r_1$ is nondeterministic.

regex $r_2 = bb^*a(b^*\mathbf{a})^*$

$\Rightarrow r_2$ is deterministic.

UNIVERSITÄT BAYREUTH

Definition (Deterministic Regular Expressions)

A regular expression is deterministic if, when reading a word from left to right, it is always clear where in the expression the current symbol can be matched without looking ahead.

Example

word $w = ba\mathbf{a}$

regex $r_1 = b(a + b)^*a$

$\Rightarrow r_1$ is nondeterministic.

regex $r_2 = bb^*a(b^*\mathbf{a})^*$

$\Rightarrow r_2$ is deterministic.

- Not every regular language has a deterministic regular expression.

  [Brüggemann-Klein et al., Inf. Comput. '98]

- Deciding if this is the case is PSPACE-complete.

  [Czerwiński et al., J. Comput. Syst. Sci. '17]
  [Groz et al., PODS '12]

UNIVERSITÄT BAYREUTH

Deterministic regular expressions

1. admit faster matching algorithms.
   [Bille et al., SODA '24]
2. enable efficient path counting algorithms.
   [Farias et al., ISWC '24] [Martens et al., VLDB '23]
3. enable efficient path enumeration algorithms.
   [Farias et al., ISWC '24] [Martens et al., VLDB '23]

} requires UFAs

	#det	#nondet
WD robotic	145,705,459	34,587
WD organic	1,337,541	5,727
DBpedia	1,609,521	2,622
Other	3,467	0

Nondet. regular expr. shapes:

- $ab \mid ac$
- variations of $a?a?a?a?$
- a^+a^+
- ...

UNIVERSITÄT BAYREUTH

- Translating to DFAs:
  - almost always no significant blowup
  - 4 expressions out of 148.7M have minimal DFAs of quadratic size:

$$(a^+|b^+)(a^+|b^+)(a^+|b^+)$$
$$(a^+|b^+)(a^+|b^+)(a^+|b^+)(a^+|b^+)$$
$$(a^+|b^+)(a^+|b^+)(a^+|b^+)(a^+|b^+)(a^+|b^+)$$
$$(a^+|b^+)(a^+|b^+)(a^+|b^+)(a^+|b^+)(a^+|b^+)(a^+|b^+)$$

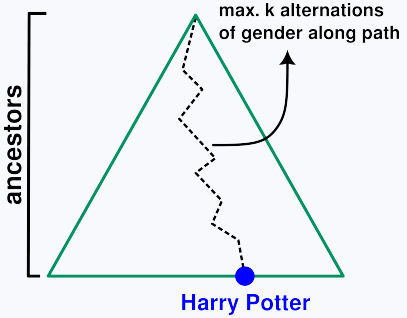
- Translating to UFAs:
  - linear-size for all expressions

UNIVERSITÄT BAYREUTH

Determinism in Regular Path Queries

Intermezzo: $(a^+|b^+)(a^+|b^+)\ldots(a^+|b^+)$

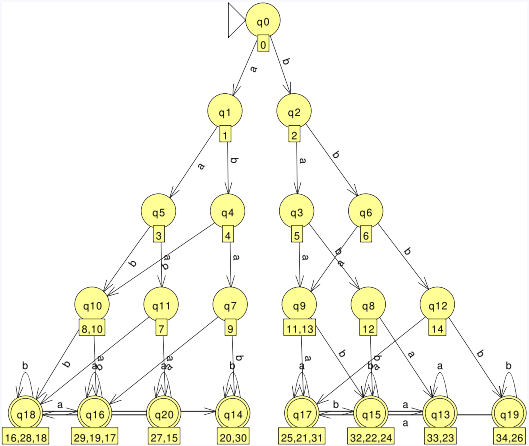
```
SELECT ?var1 ?var1Label
WHERE {
  wd:Harry_Potter
  (wdt:father+ | wdt:mother+)/
  (wdt:father+ | wdt:mother+)/
  ...
  (wdt:father+ | wdt:mother+) ?var1.


  SERVICE wikibase:label {
    bd:serviceParam wikibase:
        language "fr".
  }
}
```



UNIVERSITÄT BAYREUTH

Determinism in Regular Path Queries

Intermezzo: minimal DFA for $(a^+|b^+)(a^+|b^+)(a^+|b^+)$

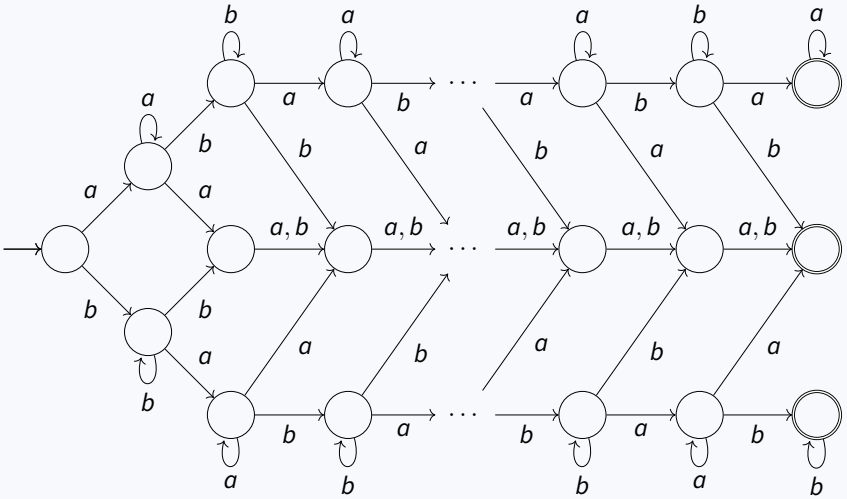


size: $\mathcal{O}(n^2)$

UNIVERSITÄT BAYREUTH

Determinism in Regular Path Queries

Intermezzo: linear-size UFA for $(a^+|b^+)(a^+|b^+)\ldots(a^+|b^+)$



size: $\mathcal{O}(n)$

UNIVERSITÄT BAYREUTH

Definition (RTQ(r))

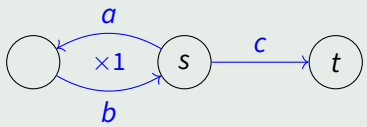
Given: A directed multigraph $G = (V, E, \text{lab})$ and $(s, t) \in V \times V$

Question: Is there a trail from s to t that matches r?

$\rightsquigarrow$ *RSPQ*(*r*) is defined analogously

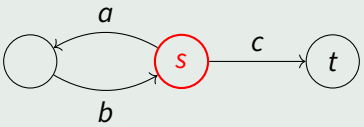
Example

RTQ($(ab)^+c$) $\rightsquigarrow$ yes



Example

RSPQ($(ab)^+c$) $\rightsquigarrow$ no



Simple Paths and Trails

T_{tract} and SP_{tract}

On directed graphs:

$$r \in T_{tract} \iff RTQ(r) \in PTIME \qquad \text{(if P} \neq \text{NP)}$$

$$r \in SP_{tract} \iff RSPQ(r) \in PTIME \qquad \text{(if P} \neq \text{NP)}$$

$$SP_{tract} \subsetneq T_{tract}$$

On undirected graphs:

$$r \in ? \iff RTQ(r) \in PTIME \qquad \text{(if P} \neq \text{NP)}$$

$$r \in ? \iff RSPQ(r) \in PTIME \qquad \text{(if P} \neq \text{NP)}$$

UNIVERSITÄT BAYREUTH

Simple Paths and Trails

T_{tract} and SP_{tract}

	RPQ $\notin T_{tract}$	Rel.	RPQ $\notin SP_{tract}$	Rel.
WD robotic	244	0.0017 ‰	226,198	1.55 ‰
WD organic	23	0.017 ‰	362	0.27 ‰
DBpedia	7	0.0043 ‰	28	0.02 ‰
Other	0	0.00 ‰	0	0.00 ‰

Example ($r \notin T_{tract}$)

- a^*ba^*
- ab^*cb^*
- $a(bc^*d)^*$
- $aa(aa)^*$

Example ($r \notin SP_{tract}$)

- a^*ba^*
- $(ab)^*$
- $a(ba)^*$
- a^*bc^*

Conclusion and Future Work

Conclusion

Analysis of 148.7M RPQs from 29 sources:

1. *Syntactical structure:*
   only 572 RPQ shapes in total
2. *Nondeterminism:*
   - Does occur, but is uncommon.
   - Linear-size DFAs for almost all RPQ shapes.
   - Linear-size UFAs for all RPQ shapes.
3. *Simple path and trail semantics:*
   Difficult RPQ shapes are uncommon.

Future Work

Investigate the interaction between RPQs and data:

- Numbers of paths that match RPQs.
- Size of compact representations.
- Evaluation under shortest path, simple path, and trail semantics *with data*.

Dataset on Zenodo

UNIVERSITÄT BAYREUTH