SHACL Constraint Validation during SPARQL Query Processing

Philipp D. Rohde
VLDB 2021 PhD Workshop
August 16th, 2021
Introduction

Knowledge Graphs (KGs) gain Momentum

Data Retrieval: SPARQL Queries

Integrity Constraints (ICs) on RDF KGs

SHACL is the standard to specify ICs on RDF KGs
Motivating Example

INPUT

KG of a University System with 37,419 entities (~1M triples)

Shape Schema: Integrity Constraints on the KG

SPARQL Query: Retrieving Data from the KG

```
SELECT ?name ?ri ?uni
WHERE {
?prof rdf:type ub:FullProfessor ;
ub:name ?name ;
ub:worksFor :Dept0 ;
ub:doctoralDegreeFrom ?uni ;
ub:emailAddress ?email ;
ub:researchInterest ?ri .
}
```

NP-complete

OUTPUT

SPARQL Query Result with SHACL Validation Result Annotation

<table>
<thead>
<tr>
<th>name</th>
<th>ri</th>
<th>uni</th>
<th><strong>meta</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>FullProfessor0</td>
<td>Research6</td>
<td><a href="http://www.University6.edu">http://www.University6.edu</a></td>
<td>all requirements met</td>
</tr>
<tr>
<td>FullProfessor3</td>
<td>Research10</td>
<td><a href="http://www.University888.edu">http://www.University888.edu</a></td>
<td>University888 violates name constraint</td>
</tr>
</tbody>
</table>

Query result annotation requires

- SPARQL query execution
- SHACL shape schema validation
- and is computationally complex
Motivating Example

**INPUT**

KG of a University System with 37,419 entities (~1M triples)

**Shape Schema:**

Integrity Constraints on the KG

**SPARQL Query:**

Retrieving Data from the KG

```
SELECT ?name ?ri ?uni
WHERE
{
?prof rdf:type ub:FullProfessor ;
ub:name ?name ;
ub:worksFor :Dept0 ;
ub:doctoralDegreeFrom ?uni ;
ub:emailAddress ?email ;
ub:researchInterest ?ri .
}
```

**OUTPUT**

SPARQL Query Result with SHACL Validation Result Annotation

Query result annotation requires:
- SPARQL query execution
- SHACL shape schema validation

and is **computationally complex**

```
name      ri       uni
FullProfessor0  Research6  http://www.Univeristy6.edu  all requirements met
FullProfessor3  Research10 http://www.University888.edu  University888 violates name constraint
...  
```
Presentation Outline

1. Related Work
2. Proposed Approach
   2.1. Problem Statement
   2.2. Proposed Solution
   2.3. Online vs Offline Validation
3. Preliminary Results
   3.1. Trav-SHACL
   3.2. SPARQL Query Result Annotation
4. Lessons Learned
5. Conclusions
Related Work

SHACL Validation

- Complexity analysis of SHACL
  - NP-complete
- Identification of tractable fragments
  - no negation, but disjunction
  - no negation in recursion
  - no recursion
- SHACL validation using SPARQL endpoints
  - before: in-memory knowledge graphs
  - Datalog-like rules
- Recursive SHACL
  - left open in the specification

Integrity Constraints in Query Processing

Abbas et al.

- well-formed ShEx schemas
- SPARQL triple pattern reordering
  - hierarchical structure of ShEx shapes
  - shapes included in other shapes are ranked higher
  - triple patterns with unique predicates are ranked highest

Rabbani et al.

- Extension of SHACL with statistics
- Cost-based query optimizer
- Precomputation time reduced

No work on explainable SPARQL query results so far.
Problem Statement

INPUT
\[ S = \langle S, \text{TARG}, \text{DEF} \rangle \]
\[ G = \langle V_G, E_G \rangle \]
\[ \Gamma_{S,G} \]
\[ Q \]

OUTPUT

SPARQL Query Execution
\[ [Q]^G \]

SHACL Schema Validation
\[ [S]^G \]

Query Result Annotation

Minimize Execution Time

Complexity
Query Evaluation: NP-c
SHACL Validation: NP-c

Finding physical plan for query Q whose execution validates the SHACL shape schema S and produces the answers of Q efficiently.
Proposed Approach

**Query Decomposition**
- subject star-shaped decomposition
- one star ≈ one class

**Query Result Annotation**
- add SHACL validation result as metadata
- explainability

**SHACL Validation**
- interleaved validation
- subset of shape schema

**Novelty of the approach:**
- identification of query plan able to combine query answering with integrity constraint validation
- explainability of SPARQL query results
- optimizations in SHACL validation
<table>
<thead>
<tr>
<th></th>
<th>online</th>
<th>offline</th>
</tr>
</thead>
<tbody>
<tr>
<td>speed</td>
<td>✗</td>
<td>✔️</td>
</tr>
<tr>
<td>who</td>
<td>✔️ everyone</td>
<td>✗ data provider</td>
</tr>
<tr>
<td>adaptivity</td>
<td>✔️</td>
<td>✗</td>
</tr>
</tbody>
</table>

Offline validation is **no option** in this scenario.
Results So Far: Trav-SHACL (1/2)


Trav-SHACL ...
- plans the traversal order
- interleaves the validation steps
- rewrites queries to make them more selective
Results So Far: Trav-SHACL (2/2)

- # Constraint query mappings:
  - 839K in SHACL2SPARQL,
  - 468K in Trav-SHACL.

- Trav-SHACL always delivers results continuously,
  - generates the first answer faster,
  - finishes the execution faster,
  - scales up to large knowledge graphs.

Impact of the interleaved execution

Metrics

dief@t: continuous efficiency at time t
(TFFF)^-1: Time for First Answer (sec)
(ET)^-1: Execution Time (sec)
Comp: sum of (in)validated entities
T: Throughput (answer/sec)
Results So Far: Query Annotation

**Single Star-Shaped Query**

```sparql
SELECT ?name ?ri ?uni
WHERE {
  ?prof rdfs:typo ub:FullProfessor ;
  ub:name ?name ;
  ub:worksFor :Dept0 ;
  ub:doctoralDegreeFrom ?uni ;
  ub:emailAddress ?email ;
  ub:researchInterest ?ri .
}
```

<table>
<thead>
<tr>
<th>Name</th>
<th>Research</th>
<th>University</th>
</tr>
</thead>
<tbody>
<tr>
<td>FullProfessor0</td>
<td>Research6</td>
<td><a href="http://www.University6.edu">http://www.University6.edu</a></td>
</tr>
<tr>
<td>FullProfessor3</td>
<td>Research10</td>
<td><a href="http://www.University888.edu">http://www.University888.edu</a></td>
</tr>
</tbody>
</table>

University888 violates name constraint

**Modified XJoin**

**RQ:** Is the performance improved by applying the proposed approach?

**WatDiv**
- 10 million triples
- 3-5 triple patterns per query
- less than 100 query results

The performance is improved but more studies are needed.
Research Plan

**SPARQL Query Result Annotation**

- extension of the semantics of SPARQL
- more studies on behavior
- extension to queries with any BGP
- integration into federated query engine

**Additional Annotations**

- verification of result via blockchain

*Extension of SPARQL for explainability. SHACL validation results only one of many possibilities.*
Lessons Learned

- Benchmark needed
- Validation computationally expensive

- Explainability needed
- Lower performance is the price

- Many KGs have low data quality
- Means to improve quality are needed
Conclusions

**Motivating Example**

**INPUT**
- KB of a University System with 37,459 entities (~5M triples)
- Shape Schema: Integrity Constraints on the KB

**SPARQL Query**
- Retrieving Data from the KB

**OUTPUT**
- SPARQL Query Result with SHACL Validation Result Annotation

**Label**
- NAME: FullProfessor
- Research: http://www.UniversityX.edu
- ResearchID: http://www.UniversityY.edu
- UniversityX fulfills none constraint

**Proposed Approach**

**Query Decomposition**
- subject viewpoint dissemination
  - one step per node class

**SHACL Validation**
- inferred validation
- subset of shape schema

**Query Result Annotation**
- add SHACL validation result as metadata
- sustainability

**Extension of SPARQL for Explainability**

**Promising Results from Prototype**

**Annotation with SHACL Validation**

**Research Plan**

**Additional Annotations**
- more studies on behavior
- extension to queries with any BGP
- formalization of a SPARQL extension
- integration into federated query engine
- verification of result via biomolecular
Thanks for your attention!

Contact:
Philipp D. Rohde
philipp.rohde@tib.eu  @philipp_rohde