A Framework to Include and Exploit Probabilistic Information in SHACL Validation Reports

Rémi Felin, **Catherine Faron** and **Andrea G. B. Tettamanzi**

ESWC 2023 - 2023/06/01

3iA Côte d'Azur
Institut interdisciplinaire d'intelligence artificielle

UNIVERSITÉ CÔTE D'AZUR

Inria

cnrs

i3s sophia antipolis

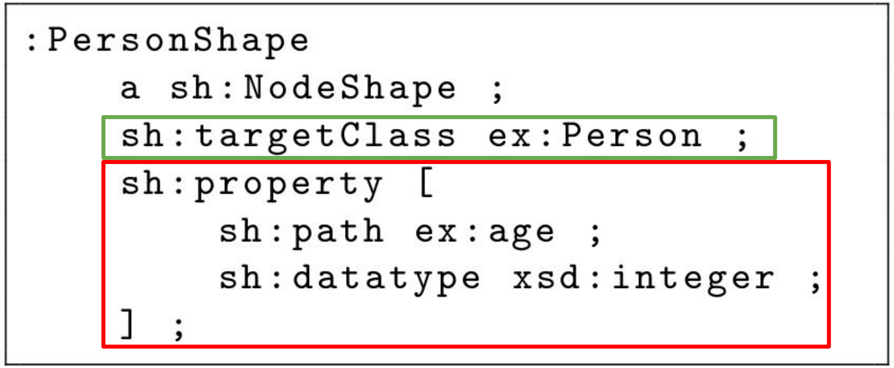
wimmics

Introduction

- **Evaluation of RDF graphs** against domain constraints
- **SHACL**, the SHApes Constraints Language
- Real-world RDF graphs are **incomplete** and contain **errors**

SHACL Shapes

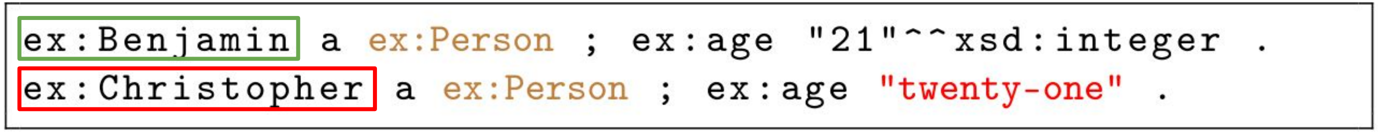
- An instance of **sh:NodeShape** or sh:PropertyShape
- targets a specific set of nodes in RDF graph
  - **sh:targetClass**
  - sh:targetNode
  - sh:targetSubjectsOf
  - ...
- evaluates these nodes against **a set of constraints**
  - value type (**sh:datatype**)
  - cardinality (sh:minCount and sh:maxCount)
  - ...



```
:PersonShape
    a sh:NodeShape ;
    sh:targetClass ex:Person ;
    sh:property [
        sh:path ex:age ;
        sh:datatype xsd:integer ;
    ] ;
```

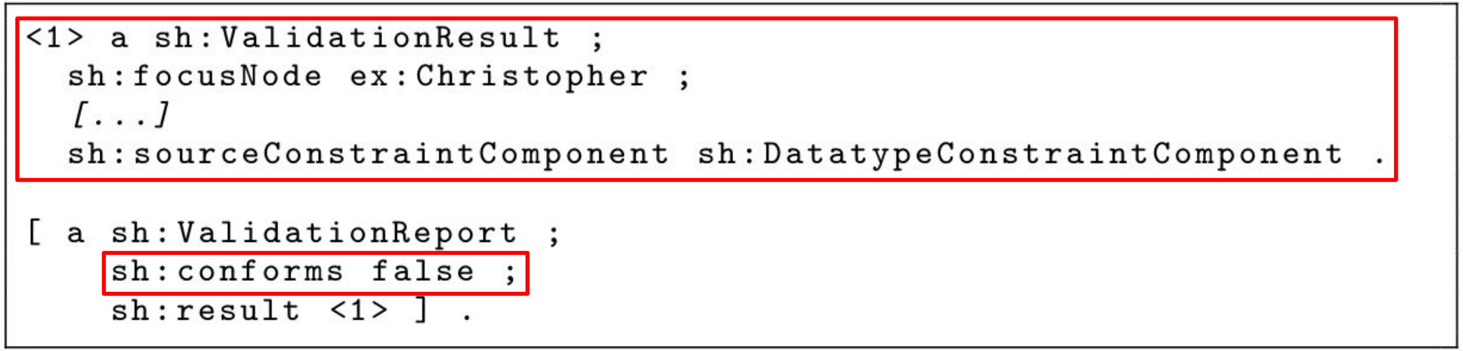
*Inspired by the SHACL shapes examples: https://www.w3.org/TR/shacl/

SHACL Validation Report



```
ex:Benjamin a ex:Person ; ex:age "21"^^xsd:integer .
ex:Christopher a ex:Person ; ex:age "twenty-one" .
```

Validate targeted nodes against the shape :PersonShape



```
<1> a sh:ValidationResult ;
  sh:focusNode ex:Christopher ;
  [...]
  sh:sourceConstraintComponent sh:DatatypeConstraintComponent .

[ a sh:ValidationReport ;
    sh:conforms false ;
    sh:result <1> ] .
```

Research Question

How to design a validation process

*considering **physiological errors** in real-life data?*

Physiological errors

In a real-world context, RDF graphs can be imperfect and incomplete

- Collaborative building of large RDF graphs (e.g. Wikidata)
- Automatically constructed RDF graphs (e.g. DBpedia)

A Probabilistic Model for SHACL Validation

Let a shape S and an RDF graph υ , we note :

- υ_S the set of **triples tested** during the validation
- υ_S^- the set of **violations**
- υ_S^+ the set of **confirmations**

$$\upsilon_S = \upsilon_S^+ \cup \upsilon_S^-$$

A Probabilistic Model for SHACL Validation

- **Assumption:** the validation process of a shape follows a *binomial distribution* considering a rate of physiological errors p

  When a triple violates a shape we consider it is a **success (1)**

  Otherwise, it is a **failure (0)**.

- **Likelihood** of observing $\|v_S^-\|$ violations in v_S

$$L_{\|v_S^-\|} = P(X = \|v_S^-\|) = \binom{\|v_S\|}{\|v_S^-\|} \cdot p^{\|v_S^-\|} \cdot (1-p)^{\|v_S^+\|}$$

A Probabilistic Model for SHACL Validation

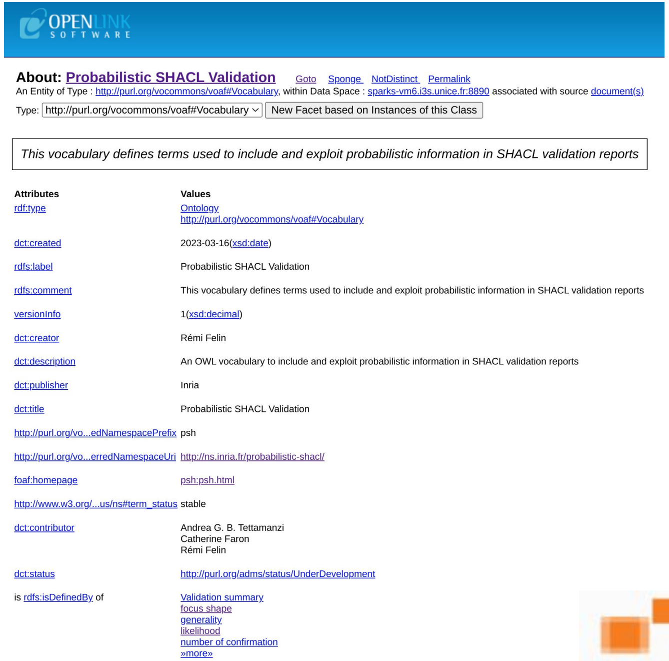
Generality measure:

$$G(S) = \frac{\|v_S\|}{\|v\|}$$

representativeness of a shape S considering v

Extended SHACL Validation Report

***Dereferencing*:**
`https://ns.inria.fr/probabilistic-shacl/`



OPENLINK SOFTWARE

About: Probabilistic SHACL Validation Goto Sponge NotDistinct Permalink

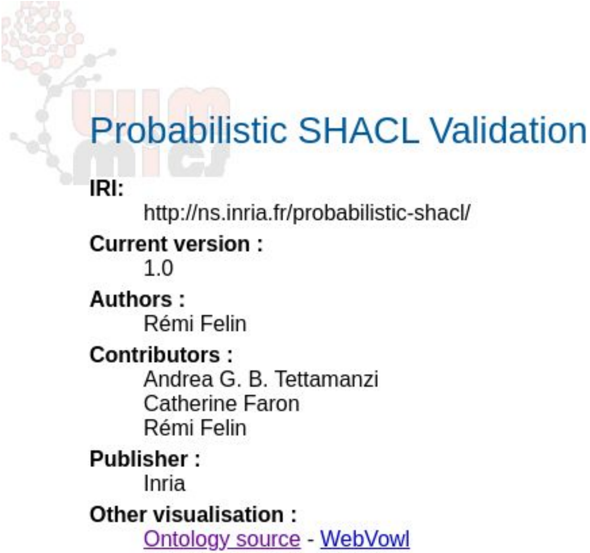
An Entity of Type : http://purl.org/vocommons/voaf#Vocabulary, within Data Space : sparks-vm6.i3s.unice.fr:8890 associated with source document(s)

Type: http://purl.org/vocommons/voaf#Vocabulary New Facet based on Instances of this Class

This vocabulary defines terms used to include and exploit probabilistic information in SHACL validation reports

Attributes	Values
rdf:type	Ontology http://purl.org/vocommons/voaf#Vocabulary
dct:created	2023-03-16(xsd:date)
rdfs:label	Probabilistic SHACL Validation
rdfs:comment	This vocabulary defines terms used to include and exploit probabilistic information in SHACL validation reports
versionInfo	1(xsd:decimal)
dct:creator	Rémi Felin
dct:description	An OWL vocabulary to include and exploit probabilistic information in SHACL validation reports
dct:publisher	Inria
dct:title	Probabilistic SHACL Validation
http://purl.org/vo...edNamespacePrefix	psh
http://purl.org/vo...erredNamespaceUri	http://ns.inria.fr/probabilistic-shacl/
foaf:homepage	psh:psh.html
http://www.w3.org/...us/ns#term_status	stable
dct:contributor	Andrea G. B. Tettamanzi Catherine Faron Rémi Felin
dct:status	http://purl.org/adms/status/UnderDevelopment
is rdfs:isDefinedBy of	Validation summary focus shape generality likelihood number of confirmation »more»

***OWL documentation*:**
`https://ns.inria.fr/probabilistic-shacl/psh.html`



Probabilistic SHACL Validation

IRI:
http://ns.inria.fr/probabilistic-shacl/

Current version :
1.0

Authors :
Rémi Felin

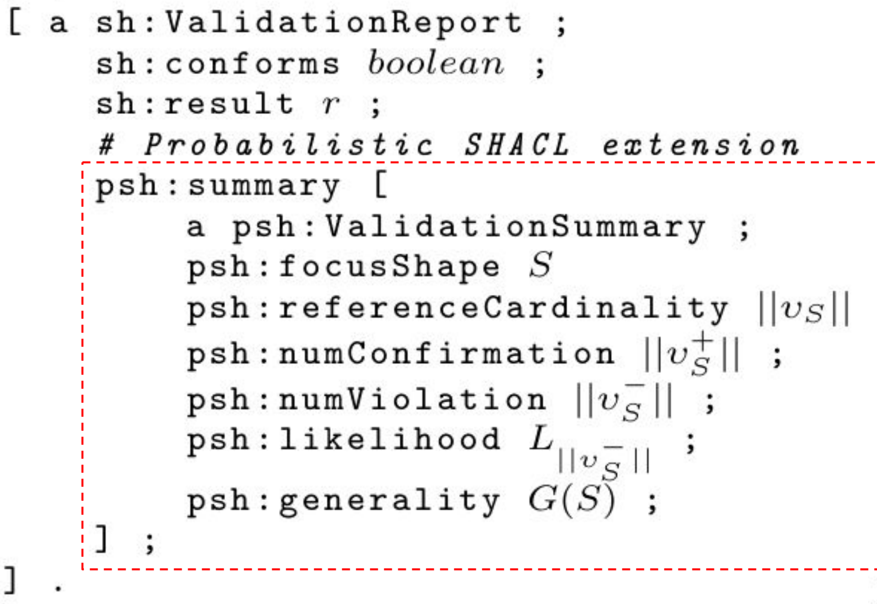
Contributors :
Andrea G. B. Tettamanzi
Catherine Faron
Rémi Felin

Publisher :
Inria

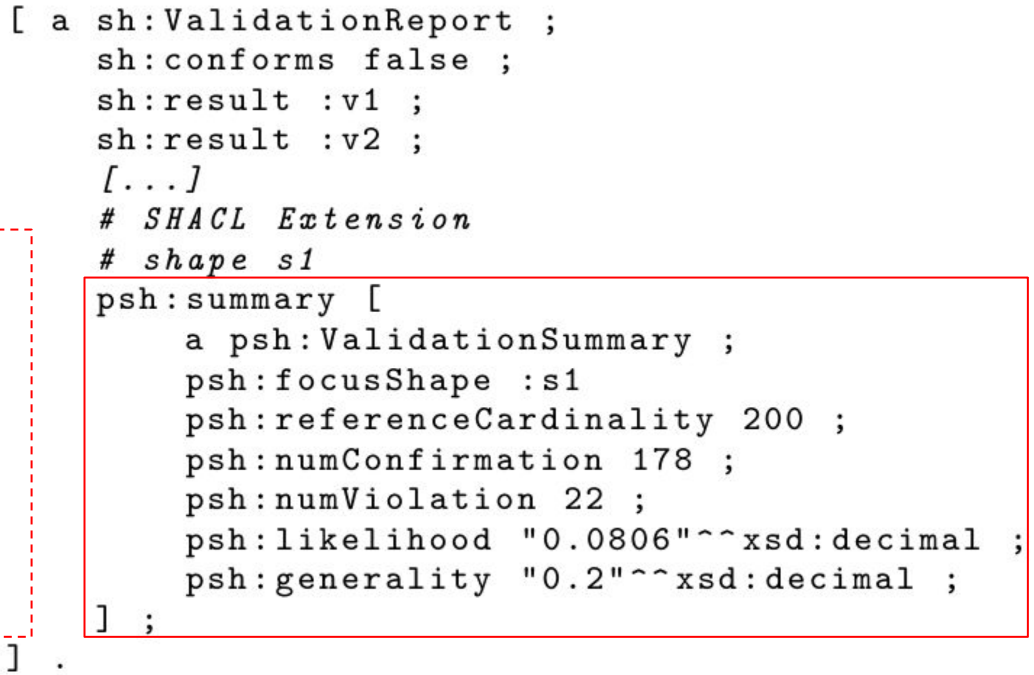
Other visualisation :
Ontology source - WebVowl

***LOV*:**
`https://lov.linkeddata.es/dataset/lov/vocabs/psh`

Extended SHACL Validation Report



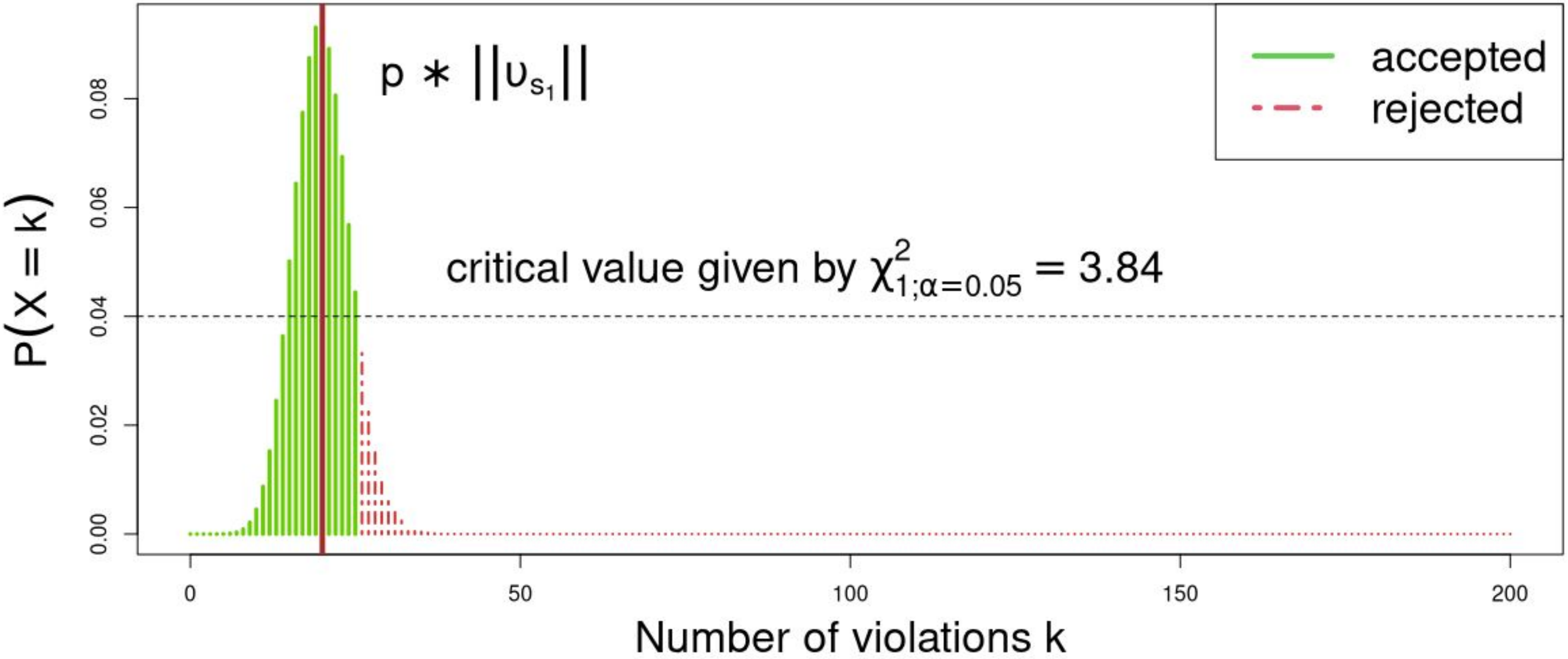
```
[ a sh:ValidationReport ;
    sh:conforms boolean ;
    sh:result r ;
    # Probabilistic SHACL extension
    psh:summary [
        a psh:ValidationSummary ;
        psh:focusShape S
        psh:referenceCardinality $||v_S||$
        psh:numConfirmation $||v_S^+||$ ;
        psh:numViolation $||v_S^-||$ ;
        psh:likelihood $L_{||v_S^-||}$ ;
        psh:generality $G(S)$ ;
    ] ;
] .
```



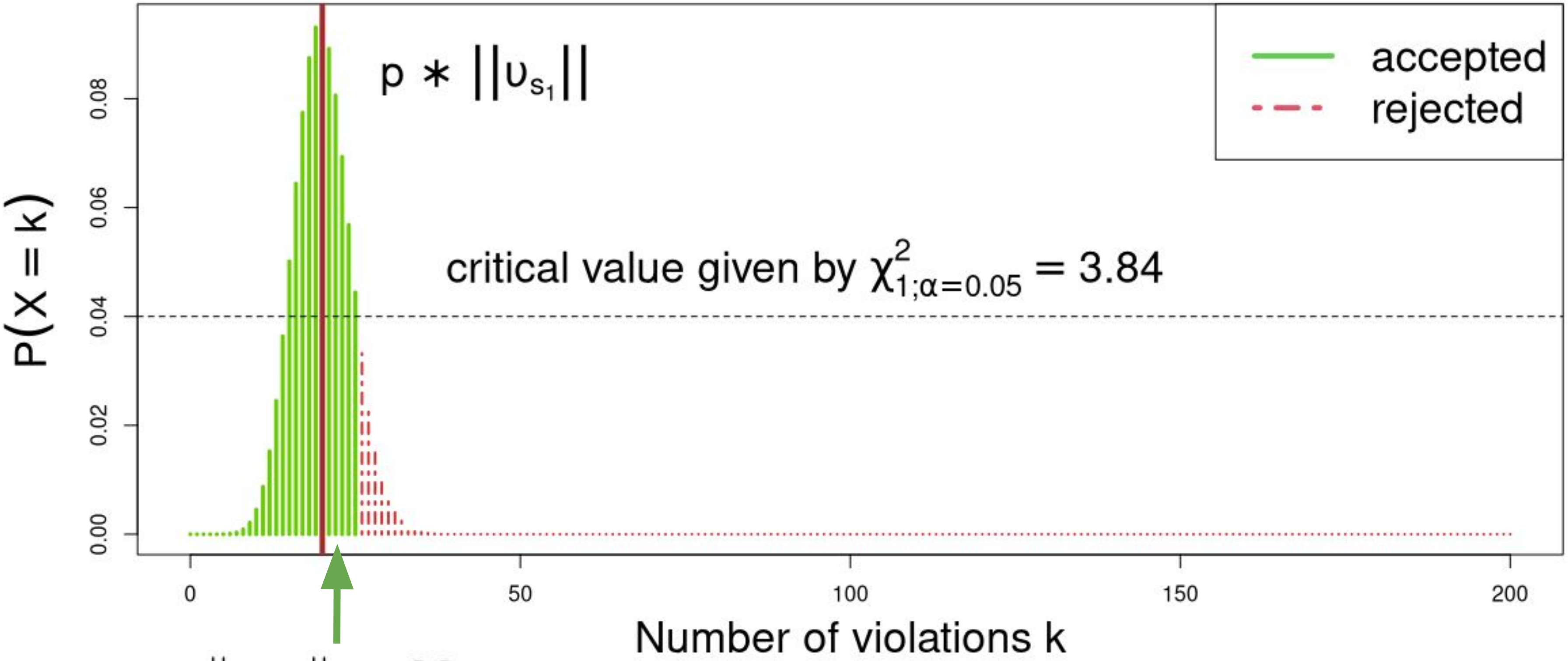
```
[ a sh:ValidationReport ;
    sh:conforms false ;
    sh:result :v1 ;
    sh:result :v2 ;
    [...]
    # SHACL Extension
    # shape s1
    psh:summary [
        a psh:ValidationSummary ;
        psh:focusShape :s1
        psh:referenceCardinality 200 ;
        psh:numConfirmation 178 ;
        psh:numViolation 22 ;
        psh:likelihood "0.0806"^^xsd:decimal ;
        psh:generality "0.2"^^xsd:decimal ;
    ] ;
] .
```

with $\|v\| = 1000$ **and** $p = 0.1$

Hypothesis Testing for Shape Acceptance

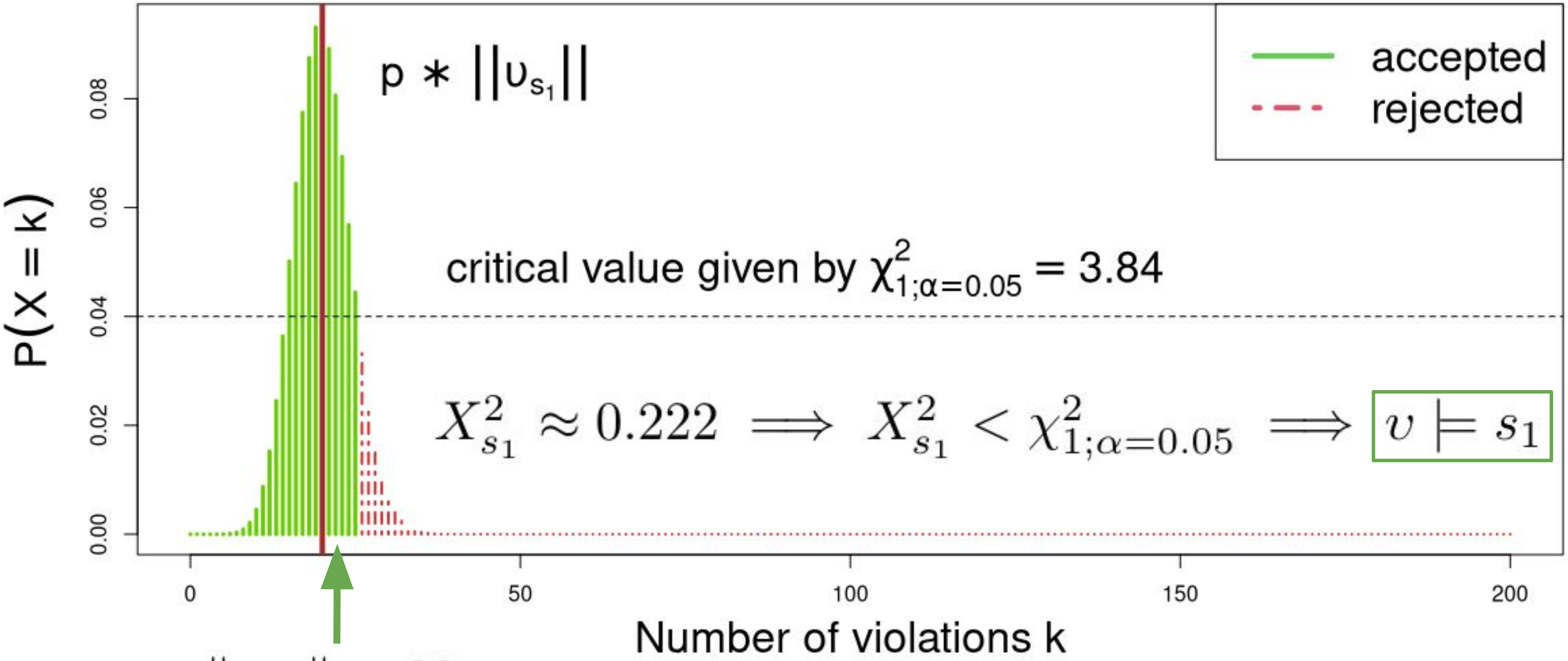


Hypothesis Testing for Shape Acceptance



$$\|\upsilon^-_{s_1}\| = 22$$

Hypothesis Testing for Shape Acceptance



$p * ||\upsilon_{s_1}||$

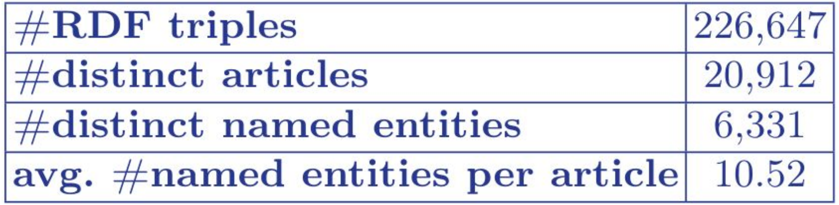
critical value given by $\chi^2_{1;\alpha=0.05} = 3.84$

$$X^2_{s_1} \approx 0.222 \implies X^2_{s_1} < \chi^2_{1;\alpha=0.05} \implies \boxed{\upsilon \models s_1}$$

$$\|\upsilon^-_{s_1}\| = 22$$

Experiments

- Evaluation of a subgraph of *CovidOnTheWeb* **[Michel & al, ISWC, 2020]** against **377 SHACL shapes**.
    - *CovidOnTheWeb*: scientific articles annotated with *Wikidata* NE



#RDF triples	226,647
#distinct articles	20,912
#distinct named entities	6,331
avg. #named entities per article	10.52

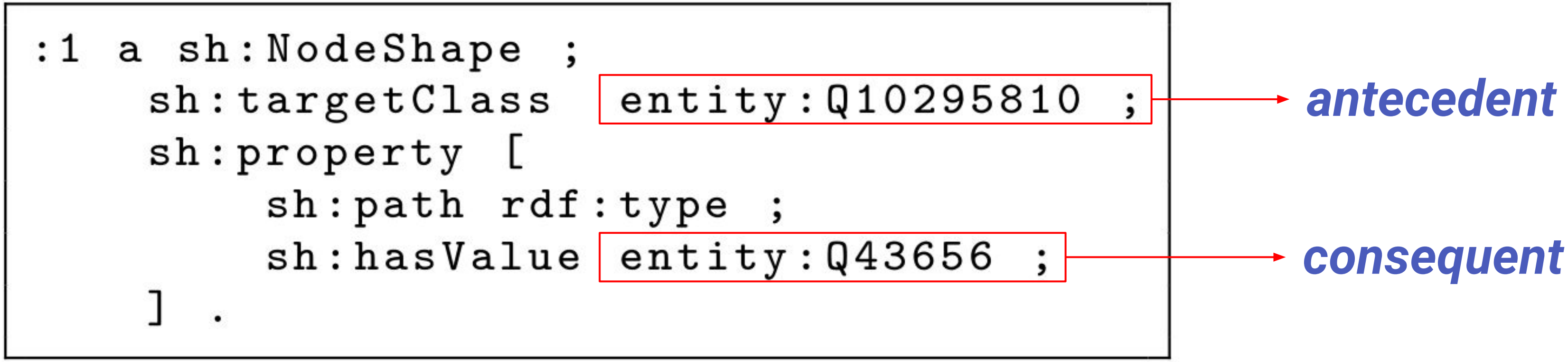
-
    - shapes represent association rules **[Cadorel & al, WI-IAT, 2020]**

- Estimation of the theoretical error proportion of the RDF graph

Evaluations performed with multiple rates of physiological errors p

Experiments

Representing association rules as SHACL shapes



```
:1 a sh:NodeShape ;
    sh:targetClass  entity:Q10295810 ;
    sh:property [
        sh:path rdf:type ;
        sh:hasValue entity:Q43656 ;
    ] .
```

antecedent

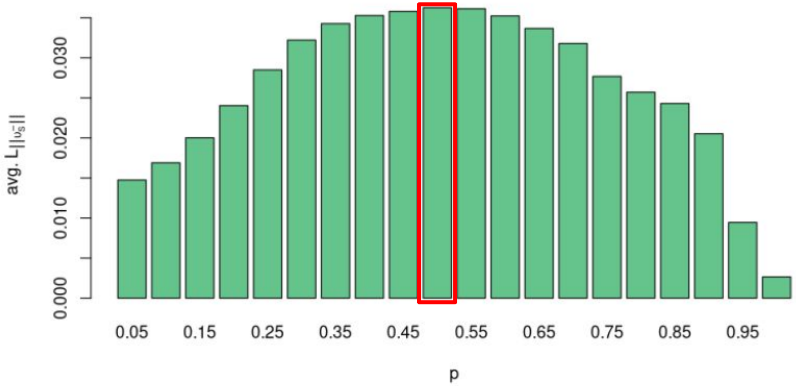
consequent

Experiments

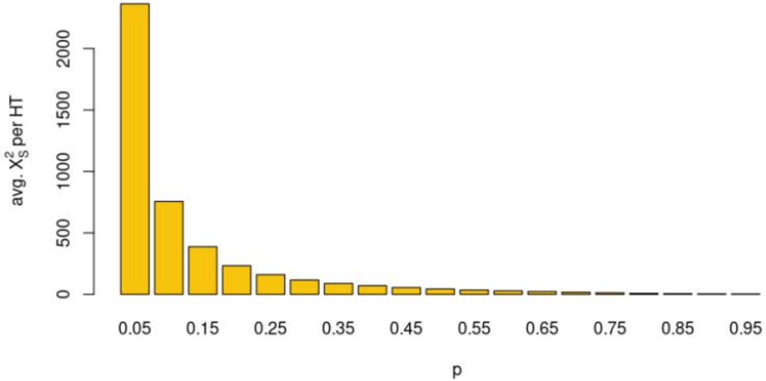
Results: Determining a Theoretical Error Proportion

Hypothesis tests performed with a **significance level** α = *5%*

$$avg(L_{\|v_S^-\|}) = 0.0362\%$$



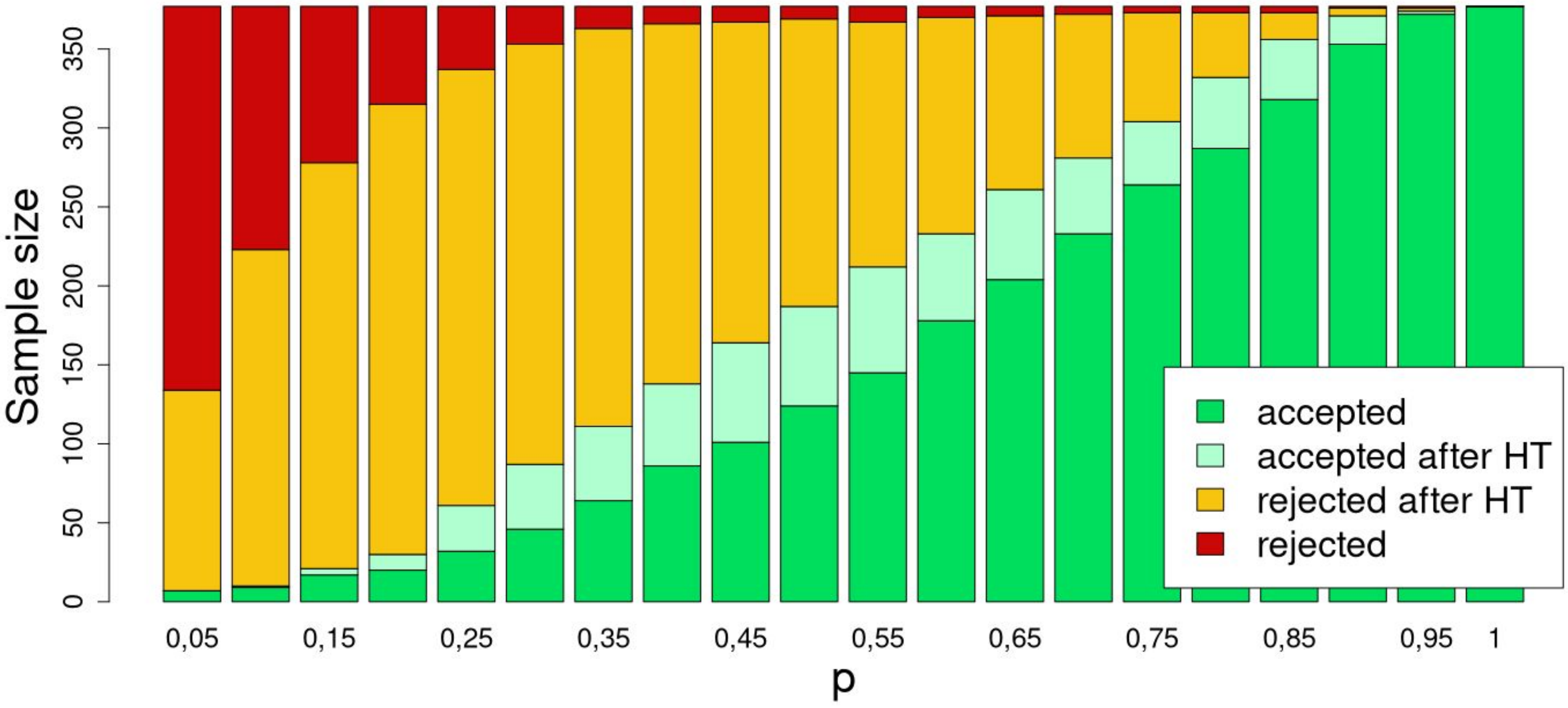
(a) $L_{||v_S^-||}$ average



(b) X_S^2 average

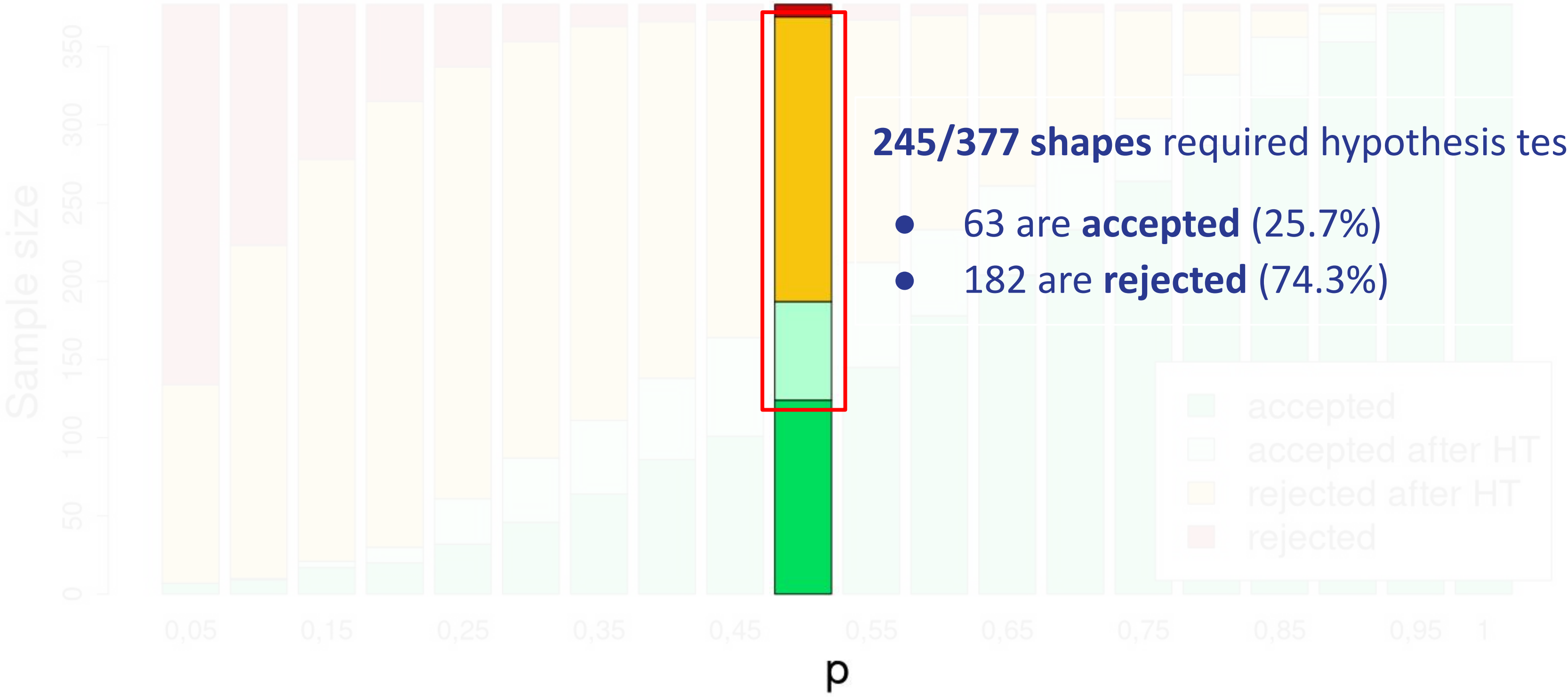
Experiments

Results: Shapes acceptance as a function of the theoretical error proportion p



Experiments

Results: Shapes acceptance as a function of the theoretical error proportion p

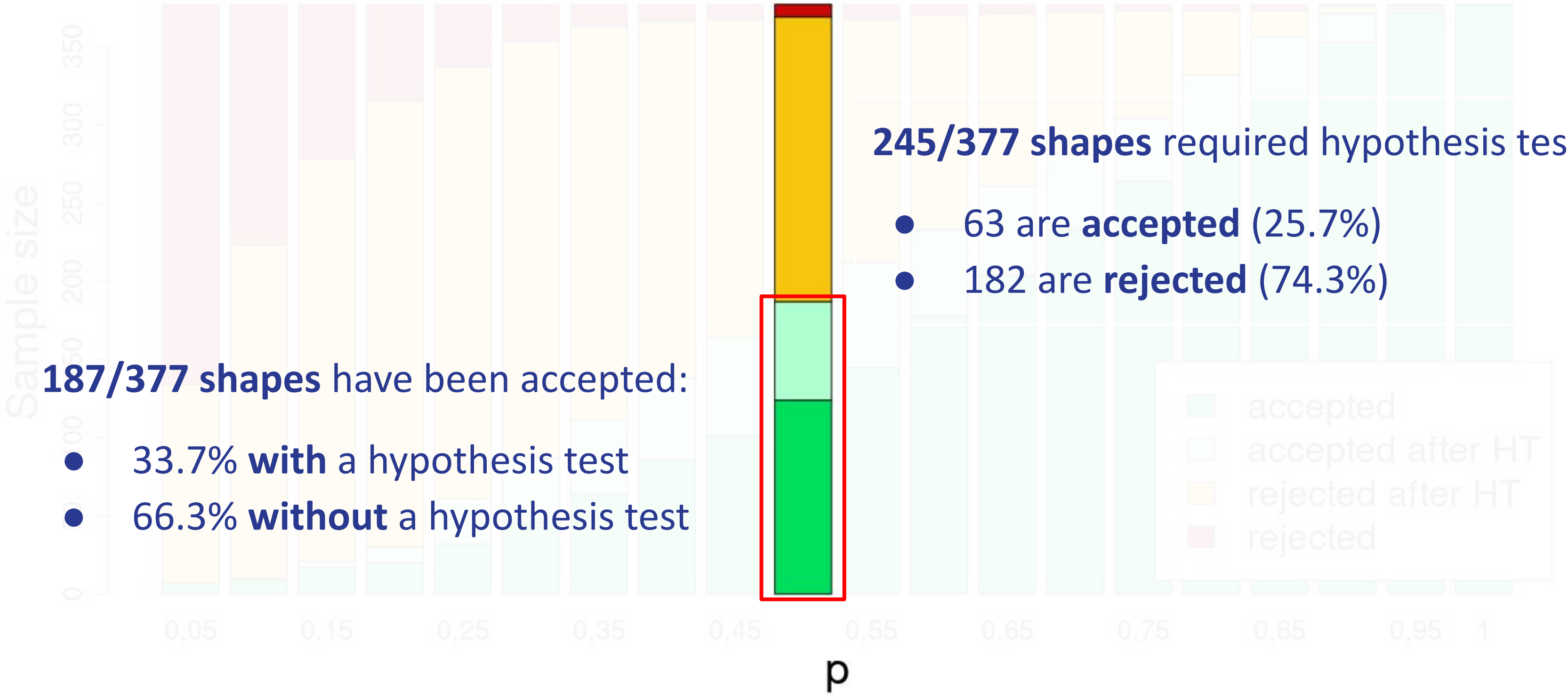


245/377 shapes required hypothesis testing:

- 63 are **accepted** (25.7%)
- 182 are **rejected** (74.3%)

Experiments

Results: Shapes acceptance as a function of the theoretical error proportion p



245/377 shapes required hypothesis testing:

- 63 are **accepted** (25.7%)
- 182 are **rejected** (74.3%)

187/377 shapes have been accepted:

- 33.7% **with** a hypothesis test
- 66.3% **without** a hypothesis test

Experiments

Results on Scalability

Computation time for the evaluation of CovidOnTheWeb against the 377 shapes:

- with standard validation: **1 minute 29**
- with a probabilistic validation: **1 minute 35**

Linear and small increase of the computation time (6.31%)

Conclusion

- A probabilistic framework relying on **likelihood** and **generality** measures
- A reliable **automatic acceptance model** based on these measures
- A model for **estimating the theoretical error proportion** from the evaluation of RDF data against a comprehensive set of SHACL shapes
- A **scalable framework** that can be applied to large RDF graphs

- Perspective: **shape mining** from RDF graphs using this probabilistic framework

Thank you !

@ remi.felin@inria.fr

in Rémi FELIN

@RemiFelin