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

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

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

Rémi Felin[1][0000−0003−2532−7555], Catherine Faron[1][0000−0001−5959−5561], and Andrea G. B. Tettamanzi[1][0000−0002−8877−4654]

Université Côte d'Azur, Inria, I3S, Sophia-Antipolis, France
{name.surname}@inria.fr

Abstract. The Shapes Constraint Language (SHACL) is a W3C recommendation which allows to represent constraints in RDF– shape graphs –, and validate RDF data graphs against these constraints. A SHACL validator produces a validation report whose result is false for a shape graph as soon as there is at least one node in the RDF data graph that does not conform to the shape. This Boolean result of the validation of an RDF data graph against an RDF shape graph is not suitable for discovering new high-potential shapes from the RDF data. In this paper, we propose a probabilistic framework to accept shapes with a realistic proportion of nodes in an RDF data graph that does not conform to it. Based on this framework, we propose an extension of the SHACL validation report to express a set of metrics including the generality and likelihood of shapes and we define a method to test a shape as a hypothesis test. Finally, we present the results of experiments conducted to validate a test RDF data graph against a set of shapes.

Keywords: RDF · SHACL · Shape Testing · Data Validation · Probabilistic Assessment

1 Introduction

The notable growth of the semantic Web has led to the emergence of new research areas such as the quality of RDF data. SHACL is the language recommended by the W3C to express patterns that RDF data must respect in order to ensure the dataset consistency.

We observe that violations generated during a SHACL validation of a shape are a significant factor. As soon as we observe at least one violation the shape is inconsistent with the RDF data. Considering a large collaborative RDF dataset with a massive and constant increase of RDF triples (e.g., DBPedia), we assume that a large number of RDF data violations against a set of shapes seems inevitable due to incomplete and/or incorrect data. In practice, a more in-depth investigation of the data seems necessary. An expert could develop a strategy for updating the data or the shapes depending on the rate or the nature of the violations. This problem has a direct impact on SHACL shape mining and limits domain knowledge learning. We tackle the following research question:

How to design a validation process considering physiological errors in real-life data?

Our contribution addresses the problem by suggesting a framework based on a probabilistic model to consider a rate of violations p assumed to be contained in an RDF dataset. p represents the proportion of errors that RDF data contains. We define a *measure of likelihood* to observe a given number of violations. We assess a given RDF dataset against a set of shapes to verify the consistency of the dataset considering a theoretical error rate.

This paper is organized as follows: In Section 2, we summarise the related work and the positioning of our work. In Section 3 we present our probabilistic model (3.1), our extension of the SHACL validation report model (3.2), and our proposal of an extended shape validation process as a test of hypothesis (3.3). We present the results of our experiments in Section 4. We conclude and discuss further research in Section 5.

2 Related Work

Given that SHACL is a fairly new recommendation, dating from 2017 [13], its interactions with other standards are subject to ongoing research. In particular, we find work on the interactions with inference rules [20], with OWL [2], description logic reasoning [15] and Ontology Design Pattern [18]. Moreover, extensions regarding SHACL validation are emerging, e.g., a SHACL validation engine based on the study of the connectivity of a given RDF graph and the collection of data in this same graph [11]. The expressiveness and semantics of SHACL is a rich subject in the literature [1, 15]: it highlights a semantics based on $\mathcal{SROIQ}$, one of the most expressive description logics.

The validation of RDF data with SHACL is a timely research question largely addressed in the literature [3, 7, 9, 12, 14, 19]. All these works consider a standard use of SHACL: an RDF dataset is valid against a shape if it verifies the expressed constraints. Our approach extends the standardized SHACL validation process to overcome its binary character by considering a possible acceptable violation rate.

SHACL constraint generation [10, 22, 23] can be carried out in several ways, some using data-based and statistical approaches, others based on ontologies [5]. The different approaches lead to different ways of tackling the validation of these shapes: The statistical-based approach requires expert analysis to define the consistency of a shape, while the ontology-based approach relies on the described RDF Schema properties (`rdfs:range`; `rdfs:domain`; ...) to provide a set of shapes based on this ontology, which can be validated if the quality of the ontology is assured. Knowledge graph profiling [21] is an important issue in order to induce constraints from large KGs [17]. The work presented in this article is focused on RDF data validation against shapes and is in line with the logic of providing expertise on the consistency of RDF data by considering the inescapable errors that they may contain, against a set of shapes that may be generated automatically or provided by an expert.

3 A Probabilistic Framework for Shape Assessment

3.1 Probabilistic Model

In a real-life context, RDF datasets are imperfect, incomplete (in the sense that expected data is missing) and containing errors of various natures. The quality control of RDF data and efficient data integration guaranteeing RDF data consistency are use cases that can be tackled using SHACL. In another respect, SHACL shapes mining from RDF data is a promising approach to learn domain knowledge (domain constraints). Candidate SHACL shapes are those triggering a few violations in the data, but this is directly correlated with the quality (error rate, which, however, is unknown) of the RDF dataset considered.

We propose to extend the evaluation of RDF data against SHACL shapes by considering a physiological theoretical error proportion p in real-life RDF data. In this context, mathematical modeling of the SHACL evaluation process, combined with an error proportion p, is based on a probabilistic model.

Definition 1. *The cardinality (or support) of a shape s, υ_s, is the set of RDF triples targeted by s and tested during the validation. We define its cardinality as the* ***reference cardinality****: $||\upsilon_s||$.*

The confirmations and violations of a shape s, respectively υ_s^+ and υ_s^-, $\upsilon_s^+ \cap \upsilon_s^- = \emptyset$, are the disjoint sets that correspond, respectively, to the triples that are consistent with s and those that violate s.

Remark 1. The sum of the number of confirmations and the number of violations of a shape s equals to the total number of triples targeted by s:

$$\upsilon_s = \upsilon_s^+ \cup \upsilon_s^- \tag{1}$$

The modelling is based on the assessment process where we define a random variable X which conceptualises a set of observations from the validation of a shape s, i.e., a set of triples υ_s; each triple $t \in \upsilon_s$ can be either a *confirmation*, $t \in \upsilon_s^+$, or a *violation*, $t \in \upsilon_s^-$.

Let us assume a single selection among υ_s for which we have two possible values: **1** if $\upsilon_1 \in \upsilon_s^-$, **0** otherwise. We conclude that a binomial distribution models this probabilistic approach, with $X \sim B(n, p)$ where $n = ||\upsilon_s||$ and p corresponds to the unavoidable theoretical error proportion, i.e. $X \sim B(||\upsilon_S||, p)$.

Definition 2. *Considering X as a random variable with the following binomial distribution $X \sim B(n, p)$ and $\Omega = \{0, 1, \ldots, n\}$, the probability to obtain exactly* ***k*** *success among* ***n*** *attempts is:*

$$\forall k \in \Omega, P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \tag{2}$$

The likelihood measure L_k determines the plausibility of obtaining k violations, i.e. $k = ||v_S^-||$, under the hypothesis of following a binomial distribution. The calculation is based on Formula 2 (see Definition 3).

Definition 3. ***The likelihood*** *to observe a number of violations* $||v_S^-||$ *among the nodes concerned by a shape* S*, i.e.* $||v_S||$*, considering* $X \sim B(||v_S||, p)$ *is*

$$L_{||v_S^-||} = P(X = ||v_S^-||) = \binom{||v_S||}{||v_S^-||} \cdot p^{||v_S^-||} \cdot (1-p)^{||v_S^+||}. \tag{3}$$

3.2 Extension of the SHACL Validation Report Model

We propose an enriched model of the SHACL validation report to express additional information for each shape considered in the report. We defined an extension to the SHACL Validation Report Vocabulary denoted by prefix `psh`.[1] For each source shape considered in the validation of an RDF graph we generate additional triples: property `psh:summary` links the validation report to a blank node of type `psh:ValidationSummary` which is the subject of several properties whose values are the result of the computation of various metrics relative to the source shape.

The focus shape is the value of property `psh:focusShape`. It is the source shape of the validation result further described in the validation summary.

The reference cardinality of a shape s $||v_s||$ is the value of property `psh:referenceCardinality` (see Definition 1).

The numbers of confirmations and violations of a shape s, respectively $||v_s^+||$ and $||v_s^-||$, are the values of properties `psh:numConfirmation` and `psh:numViolation`.

The generality $G(s) \in [0, 1]$ of a shape s measures the *representativeness* of s considering the whole RDF graph v:

$$G(s) = \frac{||v_s||}{||v||}. \tag{4}$$

It is the value of property `psh:generality`.

The likelihood of a shape s in an RDF graph v as defined in Section 3.1 is the value of property `psh:likelihood`.

Figure 1 presents an excerpt of an example validation report where:

- the SHACL shape s_1 is described by URI `:s1`;
- the cardinality of the RDF graph being validated is $||v|| = 1000$;
- the parameter of the binomial distribution is $p = 0.1$.

[1] `prefix psh: <http://ns.inria.fr/probabilistic-shacl/>`

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

Fig. 1: Structure of the extended SHACL validation report.

3.3 Data Graph Validation Against a Shape as a Hypothesis Test

The decision-making process for a given shape S is based on the probabilistic model proposed in Section 3.1, which is based on the hypothesis that a given observation follows a binomial distribution, such that $X \sim B(||\upsilon_S||, p)$. However, the question concerning the consistency of the model is relevant as it can lead to incorrect conclusions. We propose an approach based on hypothesis testing which highlights the consistency of our hypothesis and a methodology to validate our shapes.

The acceptance of a SHACL Shape s considers the proportion of violations for s, i.e $\hat{p} = \frac{||\upsilon_s^-||}{||\upsilon_s||}$. We suggest accepting the shape s as consistent with the RDF data if the observed proportion is smaller than the theoretical violation proportion:

$$\hat{p} \leq p \implies KG \models s. \tag{5}$$

In the case where the observed proportion is greater than the theoretical proportion, we minimize the distance of this probability from the maximum values of the mass function of the binomial distribution $B(||\upsilon_s||, p)$ by using hypothesis testing. Figure 3 shows the proportion of the number of violations that we accept compared to the number that we reject with our method.

The Null and Alternate Hypothesis are (respectively) H_0: *data follow the given distribution*, i.e. the frequency of observed violations $\hat{p} = \frac{||\upsilon_S^-||}{||\upsilon_S||}$ is in line with the expected proportions of violations p and $X \sim B(||\upsilon_S||, p)$. Finally, H_1 indicates that *data do not follow the given distribution.*

```
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix psh: <http://ns.inria.fr/probabilistic-shacl/> .
@prefix : <http://www.example.com/myDataGraph#> .

# SHACL Standard
:v1 a sh:ValidationResult ;
    sh:focusNode :n1 ;
    [...]
    sh:sourceShape :s1 .

:v2 a sh:ValidationResult ;
    sh:focusNode :n2 ;
    [...]
    sh:sourceShape :s1 .

[...]

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

Fig. 2: Example of an extended SHACL validation report for a shape `:s1` with $||v|| = 1000$ and $p = 0.1$.

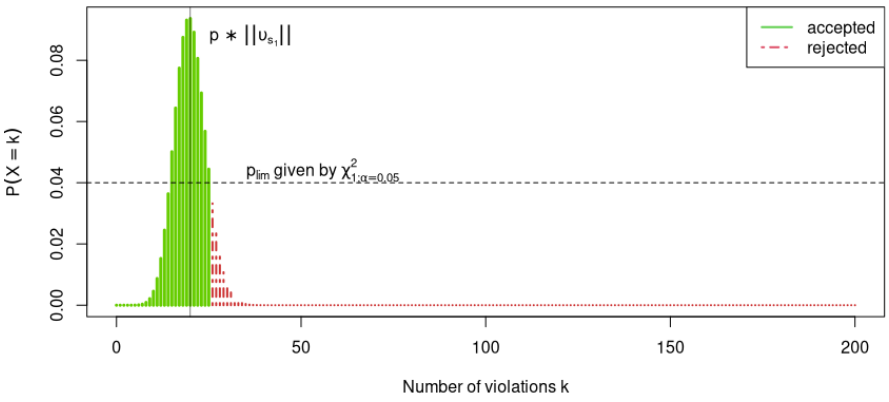


Fig. 3: Acceptance zone of shape s_1, considering $X \sim B(||\upsilon_{s_1}||, p)$ where $||\upsilon_{s_1}|| = 200$ and $p = 0.1$.

The testing for Goodness of Fit verifies the alignment of our observations with a theoretical distribution: we define X_s^2 the **test statistic** for a shape s which follows $\chi^2_{k-1,\alpha}$ assuming H_0, i.e. $X_s^2 \sim \chi^2_{k-1,\alpha}$ (a chi-square distribution with $k-1$ degrees of freedom and a level of significance $1-\alpha$) if X_s^2 verifies Definition 4. This test is performed at the α level defined at 5%. It considers k the total number of groups, i.e. $k = 2$, n_i the observed number of individuals and T_i the theoretical number of individuals. The test statistic X_s^2 is defined by

$$X_s^2 = \sum_{i=1}^{k} \frac{(n_i - T_i)^2}{T_i} \sim \chi^2_{k-1;\alpha}. \tag{6}$$

Remark 2. A shape s for which we observe very small support $||\upsilon_s||$ (let us say $||\upsilon_s|| = 5$) implies a proportion of violations and/or confirmations that are less than 5. Thus, the hypothesis test cannot be applied because the sample is not sufficiently representative of a Chi-square distribution (see Definition 4).

Definition 4. *The testing for goodness of fit is applicable (Formula 6) if* $\forall i \in [1, k], T_i \geq 5$.

The critical region i.e. the rejection region of H_0, is defined by the value $\chi^2_{k-1;\alpha}$. Considering $\alpha = 0.05$ and $k = 2$, we define the critical value: $\chi^2_{k-1;\alpha} = \chi^2_{1;\alpha=0.05} = 3.84$.

Remark 3. An alternative formula considers the acceptance interval I_a of a χ^2 distribution, i.e. $I_a = [0, \chi^2_{k-1;\alpha}]$ which accept H_0 if $X_s^2 \in I_a$.

The acceptance of the null hypothesis , i.e., $X \sim B(||\upsilon_s||, p)$, implies that the value of our test statistic X_s^2 is not included in the rejection zone of the $\chi^2_{k=1}$ distribution, such that

$$X_s^2 \leq \chi^2_{k-1;\alpha}. \tag{7}$$

The acceptance of H_0 implies the acceptance of the considered shape s, i.e.,

$$X_s^2 \leq \chi^2_{k-1;\alpha} \implies KG \models s. \tag{8}$$

Let us consider the case shown in Figure 2 as an example of an application. We observe a proportion of violations that is slightly higher than expected, i.e., $\hat{p} = \frac{||\upsilon_{s_1}^-||}{||\upsilon_{s_1}||} = 0.11$ and $\hat{p} > p$: an analysis through the hypothesis test determines if this observation is inconsistent with the null hypothesis, and in which case we would reject H_0 and the shape `:s1`. We assume $\alpha = 5\%$ to assess $X_{s_1}^2$:

$$X_{s_1}^2 = \frac{(22-20)^2}{20} + \frac{(178-180)^2}{180} = \frac{4}{20} + \frac{4}{180} \approx 0.222.$$

The statistical test demonstrated that $X_{s_1}^2 \leq \chi^2_{1;\alpha=0.05}$ (i.e. 3.84) and so $X_{s_1}^2 \in I_a$. We accept H_0 and validate the adequacy of this hypothesis, i.e. the assumption that our observations from the validation of `:s1` follow a binomial distribution $X \sim B(200, 0.1)$, with a level of significance of $1 - \alpha$, i.e., 95%.

4 Experiments

These contributions lead to an extension of the validation report to cover the generation of a degree of probability expressed under the hypothesis that the samples follow a binomial distribution with a cardinality defined by the SHACL shapes (i.e. $||\upsilon_s||$) and a probability p defined *empirically* corresponding to the assumed proportion of violations that we accept from some RDF data. At the same time, we investigate whether such an approach can capture the knowledge domain in a larger way, i.e., a broader spectrum of accepted shapes for which they are considered consistent despite the observed violations. Considering a shape graph representative of an RDF dataset, the conclusion of an error rate p for which it is reasonable to consider the acceptance of shapes on a subset of the global dataset seems a relevant perspective for the evaluation of this work. This implies a detailed analysis of the characteristics of the considered subset, the proportions of accepted or rejected shapes and the impact of hypothesis testing on acceptance.

4.1 Experimental Setup

Our experiments use the *CovidOnTheWeb dataset*[2] [16] against **a set of 377 shapes** from a translation of the experimental results of Cadorel & al. [4] which are considered as **representative shapes** of the whole CovidOnTheWeb

[2] https://github.com/Wimmics/CovidOnTheWeb

dataset. We run the probabilistic SHACL validation engine (see Section 3.2) implemented in the *Corese* semantic web factory. We will conduct an analysis of the theoretical error rate in order to find an optimal rate: we assume the values of p empirically such that $p \in \{0.05, 0.1, 0.15, \ldots, 0.95, 1\}$, which gives 20 values for p to be tested. The experiments were performed on a Dell Precision 3561 equipped with an Intel(R) 11th Gen Core i7-11850H processor, with 32 GB of RAM running under the Fedora Linux 35 operating system. The source code is available in a public repository.[3]

CovidOnTheWeb is an RDF knowledge graphs produced from *COVID-19 Open Research Dataset (CORD-19)*. It targets **articles**, described by URIs and **named entities** identified in these articles, disambiguated by *Entity-Fishing* and linked to *Wikidata* entities. Figure 4 shows an excerpt of RDF description in CovidOnTheWeb in *turtle* format and Table 1 shows the characteristics of the RDF dataset. We consider a subset containing approximately 18.79% of the articles and 0.01% of the named entities.

Table 1: Summary of the *CovidOnTheWeb* RDF subgraph considered for the experiment.

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

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix covid:    <http://ns.inria.fr/covid19/> .
@prefix entity:   <http://www.wikidata.org/entity/> .

covid:ec1[...]2c5   rdf:type    entity:Q4407  .
covid:fff[...]86d   rdf:type    entity:Q10876 .
[...]
entity:Q4407        rdfs:label  "methyl"@en .
entity:Q10876       rdfs:label  "bacteria"@en .
```

Fig. 4: Example of RDF data extracted from *CovidOnTheWeb*.

[3] https://github.com/RemiFELIN/RDFMining/tree/eswc_2023

The candidate shapes describe association rules obtained by Cadorel & al. [4] from a subset of the *CovidOnTheWeb dataset*. These rules are not necessarily perfect, so we are interested in using them in our probabilistic approach. From the experimental results of Cadorel & al., we extracted the named entities corresponding to the antecedents and the consequents of these association rules. We have carried out a treatment allowing the conversion of these rules into SHACL shapes. We target articles belonging to a named entity, representing the *antecedent*, with the property `sh:targetClass`. Among the articles considered, we are interested in determining the affiliation to another named entity, representing the *consequent*: we use a constraint applied on the articles' type and target a named entity with the property `sh:hasValue`. In this context, a violation will invoke a violation of type `sh:HasValueConstraintComponent` for the current shape. An example of a shape formed after treatment is shown in Figure 5.

```
@prefix : <http://www.example.com/myDataGraph#> .
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix entity: <http://www.wikidata.org/entity/> .

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

Fig. 5: Example SHACL shape representing an association rule with `entity:Q10295810` (`"hypocholesterolemia"@en`) as an *antecedent* and `entity:Q43656` (`"cholesterol"@en`) as a *consequent*.

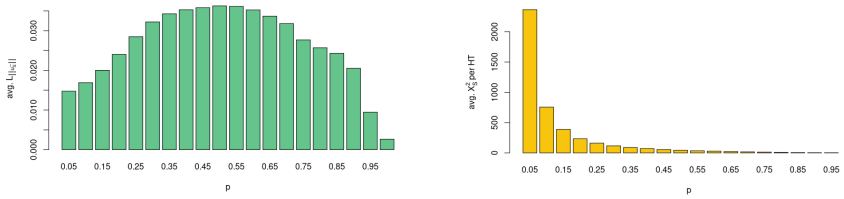
4.2 Results

Table 2 shows the first experimental results, notably the generality score which is relatively low, indicating a low average cardinality compared to the number of total triples in our dataset: approximately 106 RDF triples on average are targeted by our shapes (0.047% of the RDF triples). The rate of violations is relatively high but is nuanced by the rate of confirmations (33.19%). It highlights the interest in a probabilistic approach in order to check the consistency of our RDF dataset against the shape graph considering varying p error rates and understand how we can consider a reasonable error rate and a consistent number of valid shapes.

Figure 6a shows an increasing evolution of the likelihood measure up to the value $p = 0.5$ and then a decrease. It appears that the most reasonable error rate is **50%**, as it maximises the mean likelihood value (0.0362%).

Table 2: Summary of the SHACL shape graph considered in the experiment.

#named entities represented	337 (*5.32%*)
avg. reference cardinality	106.69 (*0.0470%*)
avg. #confirmations	33.19 (*31.11%*)
avg. #violations	73.50 (*70.89%*)
avg. generality $G(S)$ (Formula 4)	0.0005%



(a) $L_{||v_S^-||}$ average (b) X_S^2 average

Fig. 6: Average value of (a) likelihood measures and (b) statistic test as functions of the theoretical error proportion p.

Figure 7 presents the set of decisions made on shapes (acceptance, rejection) as a function of the theoretical error proportion p and clearly shows the importance of hypothesis testing. The number of tests performed increases until $p = 0.3$ and then decreases. Similarly, hypothesis testing tends to reject shapes for "small" values of p and the trend reverses as p increases: the number of accepted shapes increases and the value of the test statistic decreases (see Figure 6b). Further analysis of the results obtained with $p = 0.5$ shows that 63 shapes among the 187 accepted shapes are accepted after performing a hypothesis test, i.e. 33.7% of the accepted shapes. These same tests accepted 25.7% of the shapes that were tested, which shows its ability to efficiently filter with a risk of $\alpha = 0.05$ or 5% of being incorrect.

The production of the results in HTML format was performed with a STTL transformation [6]. STTL is an extension to the SPARQL query language to transform RDF in any template-specified text result format, which is populated with the results of a SPARQL query. In our case, we provided an HTML template including the desired values in its structure. An excerpt of 20 out of 377 results obtained for a theoretical error proportion of $p = 0.5$ is presented in Figure 8.

We compared the computation time of our proposed probabilistic validation framework with that of standard validation. For our base of 377 shapes and our extract of *CovidOnTheWeb* (226,647 triples), we observed an overall computation time of **1 minute 35 seconds** for the probabilistic validation framework against **1 minute 29 seconds** for standard validation: the probabilistic framework takes **6,31%** more time than standard validation and it is linear which makes it practical and scalable.

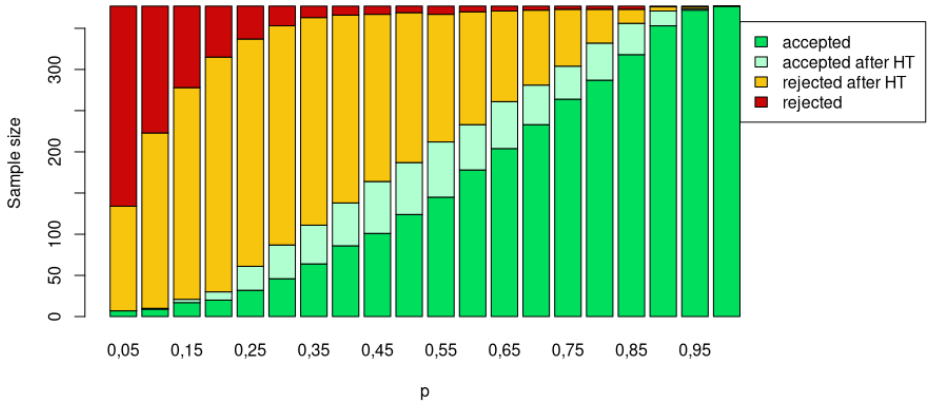


Fig. 7: Shapes acceptance as a function of the theoretical error proportion p (HT= Hypothesis Testing).

antecedent	consequent	referenceCardinality	#violation	likelihood	generality	X^2_S	Acceptance
two-hybrid screening	protein–protein interaction	48	19	0.041004880900459284	0.00021178308117910231		true
nidovirales	proteolysis	80	69	8.6669313322632E-12	0.0003529718019651705 3	42.05	false
intensive care medicine	acute respiratory distress syndrome	166	139	9.193409214822706E-20	0.0007324164890777288	75.56626506024097	false
astrocyte	central nervous system	70	34	0.09238587705330051	0.0003088503267195242		true
dopamine	serotonin	10	6	0.205078125	0.00004412147524564632	0.4	true
crystallography	crystal structure	20	7	0.0739288330078125	0.00008824295049129263		true
human parainfluenza	adenoviridae	237	133	0.00880821375320367	0.0010456789633218177	3.548523206751055	true
carbohydrate	lectin	114	75	2.4200572197826046E-4	0.000502984817800368	11.368421052631579	false
mycoplasma bovis	bovine coronavirus	12	6	0.2255859375	0.00005294577029477558		true
crystallization	diffraction	31	21	0.020653086248785257	0.0001367765732615035 8	3.903225806451613	false
membrane raft	methyl	32	19	0.080879218876361 85	0.0001411887207860682	1.125	true
ifitm1	ifitm3	27	9	0.03491956740617752	0.0001191279831632 4504		true
multiple sclerosis	myelin	139	97	1.0209205741082355E-6	0.0006132885059144837	21.762589928057555	false
wheeze	asthma	85	44	0.08188889187584301	0.0003750325395879 9367	0.10588235294117647	true
influenza a virus subtype h5n1	avian influenza	277	165	2.9696484716868 76E-4	0.001222164864304403	10.1407942238267 15	false
hepatocellular carcinoma	liver cirrhosis	72	46	0.00584315589512 9734	0.0003176746217686 5343	5.555555555555555	false
diffraction	x-ray crystallography	16	7	0.174560546875	0.0000705943603930341		true
feline infectious peritonitis	feline coronavirus	130	46	2.6051939133257 92E-4	0.000573579178193402		true
aedes aegypti	culicidae	21	4	0.0028538703918457 03	0.0000926550980158572 6		true
monomer	oligomer	83	70	5.4692741602999564E-11	0.0003662082445388644	39.144578313253014	false

Fig. 8: SHACL validation report in HTML format for $p = 0.5$.

5 Conclusion

In this article, we propose a probabilistic framework for SHACL validation, thus contributing to RDF data quality control. We extend the SHACL validation report to express the likelihood measure for the number of violations observed and we propose a decision model for a probabilistic acceptance of RDF triples against SHACL shapes. Our experiments show the capabilities of our approach to validate a real-world RDF dataset against a set of SHACL shapes while accepting a reasonable error rate p. As future work, we plan to extend our proposed framework to complex shapes, especially recursive shapes which are the focus of ongoing research [3, 8, 19]. We also plan to investigate the automatic extraction or generation of SHACL shapes from reference RDF datasets, to capture domain knowledge as constraints.

Acknowledgements This work has been partially founded by the 3IA Côte d'Azur "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

References

1. Bogaerts, B., Jakubowski, M., den Bussche, J.V.: Expressiveness of shacl features. In: ICDT (2022)
2. Bogaerts, B., Jakubowski, M., Van den Bussche, J.: Shacl: A description logic in disguise (08 2021)
3. Boneva, I., Labra Gayo, J.G., Prud 'Hommeaux, E.G.: Semantics and Validation of Shapes Schemas for RDF. In: ISWC2017 - 16th International semantic web conference. Vienna, Austria (Oct 2017)
4. Cadorel, L., Tettamanzi, A.: Mining rdf data of covid-19 scientific literature for interesting association rules. 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) pp. 145–152 (2020)
5. Cimmino, A., Fernández-Izquierdo, A., García-Castro, R.: Astrea: Automatic generation of shacl shapes from ontologies. In: Harth, A., Kirrane, S., Ngonga Ngomo, A.C., Paulheim, H., Rula, A., Gentile, A.L., Haase, P., Cochez, M. (eds.) The Semantic Web. pp. 497–513. Springer International Publishing, Cham (2020)
6. Corby, O., Faron Zucker, C.: STTL: A SPARQL-based Transformation Language for RDF. In: 11th International Conference on Web Information Systems and Technologies. Lisbon, Portugal (May 2015)
7. Corman, J., Florenzano, F., Reutter, J.L., Savkovic, O.: Validating shacl constraints over a sparql endpoint. In: International Workshop on the Semantic Web (2019)
8. Corman, J., Reutter, J.L., Savković, O.: Semantics and validation of recursive shacl. In: Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M.C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.A., Simperl, E. (eds.) The Semantic Web – ISWC 2018. pp. 318–336. Springer International Publishing, Cham (2018)
9. Debruyne, C., McGlinn, K.: Reusable shacl constraint components for validating geospatial linked data (short paper). In: GeoLD@ESWC (2021)

10. Fernandez-Álvarez, D., Labra-Gayo, J.E., Gayo-Avello, D.: Automatic extraction of shapes using shexer. Knowledge-Based Systems **238**, 107975 (2022). https://doi.org/https://doi.org/10.1016/j.knosys.2021.107975, https://www.sciencedirect.com/science/article/pii/S0950705121010972
11. Figuera, M., Rohde, P.D., Vidal, M.E.: Trav-shacl: Efficiently validating networks of shacl constraints. In: Proceedings of the Web Conference 2021. p. 3337–3348. WWW '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442381.3449877
12. K Soman, R.: Modelling construction scheduling constraints using shapes constraint language (shacl). pp. 351–358 (07 2019). https://doi.org/10.35490/EC3.2019.170
13. Kontokostas, D., Knublauch, H.: Shapes constraint language (SHACL). W3C recommendation, W3C (Jul 2017), https://www.w3.org/TR/2017/REC-shacl-20170720/
14. Köcher, A., Vieira da Silva, L.M., Fay, A.: Constraint checking of skills using shacl (07 2021)
15. Leinberger, M., Seifer, P., Rienstra, T., Lämmel, R., Staab, S.: Deciding shacl shape containment through description logics reasoning. In: Pan, J.Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web – ISWC 2020. pp. 366–383. Springer International Publishing, Cham (2020)
16. Michel, F., Gandon, F., Ah-Kane, V., Bobasheva, A., Cabrio, E., Corby, O., Gazzotti, R., Giboin, A., Marro, S., Mayer, T., Simon, M., Villata, S., Winckler, M.: Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research. In: ISWC 2020 - 19th International Semantic Web Conference. Athens / Virtual, Greece (Nov 2020). https://doi.org/10.1007/978-3-030-62466-8_19
17. Mihindukulasooriya, N., Rashid, M.R.A., Rizzo, G., García-Castro, R., Corcho, O., Torchiano, M.: Rdf shape induction using knowledge base profiling. p. 1952–1959. SAC '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3167132.3167341, https://doi.org/10.1145/3167132.3167341
18. Pandit, H., O'Sullivan, D., Lewis, D.: Using ontology design patterns to define shacl shapes. In: WOP@ISWC. pp. 67–71. Monterey California, USA (2018)
19. Pareti, P., Konstantinidis, G.: A review of shacl: From data validation to schema reasoning for rdf graphs. In: Reasoning Web (2021)
20. Pareti, P., Konstantinidis, G., Norman, T.J., Şensoy, M.: Shacl constraints with inference rules. In: Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., Gandon, F. (eds.) The Semantic Web – ISWC 2019. pp. 539–557. Springer International Publishing, Cham (2019)
21. Principe, R., Maurino, A., Palmonari, M., Ciavotta, M., Spahiu, B.: Abstat-hd: a scalable tool for profiling very large knowledge graphs. The VLDB Journal **31** (09 2021). https://doi.org/10.1007/s00778-021-00704-2
22. Rabbani, K., Lissandrini, M., Hose, K.: Shacl and shex in the wild: A community survey on validating shapes generation and adoption. In: Companion Proceedings of the Web Conference 2022. p. 260–263. WWW '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3487553.3524253
23. Wright, J., Rodríguez Méndez, S.J., Haller, A., Taylor, K., Omran, P.G.: Schímatos: A shacl-based web-form generator for knowledge graph editing. In: Pan, J.Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web – ISWC 2020. pp. 65–80. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-62466-8_5