Towards Reconciling Certain Answers and SPARQL: Bag Semantics to the Rescue?*

Sebastian Skritek

TU Wien, skritek@dbai.tuwien.ac.at

1 Introduction

Despite the intention of RDF, the data model for the Semantic Web, to support reasoning about RDF data (as witnessed e.g. by three different semantics for RDF – simple/RDF/RDFS entailment – or accompanying standards like OWL), SPARQL, the standardized query language for RDF, lacks some support of including such reasoning into query answering, as was noted e.g. in [6, 1].

One major cause of this is that SPARQL only makes partially use of *certain answers*, which are the semantics commonly applied in reasoning scenarios. While not fully supporting certain answers might be a reasonable decision when looking at the costs of evaluating and (for a user) understanding a query, it negatively affects the power of SPARQL. As a result, in recent years suggestions have been made how to cover different aspects of reasoning about RDF data (e.g. blank nodes under RDF simple entailment [6]; OWL reasoning [1]) in SPARQL by adopting a certain answer semantics.

However, as observed in [1], just applying the usual definition of certain answers to SPARQL can, in certain cases, be unnecessarily restrictive. To better illustrate these situations, let us first recall the definition of certain answers.

Given an RDF graph G (possibly extended by some additional information), most reasoning formalisms define the semantics of G in terms of an (infinite) set $models(G)$ of RDF graphs implied by G. Query answering in such a setting is then commonly defined in terms of the *certain answer semantic*: for a query Q, the certain answers $certain(Q, G)$ (w.r.t. some reasoning formalism) are

$$certain(Q, G) = \bigcap_{G' \in models(G)} Q(G'),$$

where $Q(G')$ denotes the result of evaluating Q over the RDF graph G'.

While this definition can be immediately applied to SPARQL queries, problems arise e.g. when looking at the `OPTIONAL` operator, or more precisely at the weakly monotone classes of SPARQL queries, like the well-designed queries [10].

Example 1. Consider the SPARQL query Q

`SELECT` ?`auth`, ?`award` `WHERE` {?`auth` `writes` ?`b`} `OPTIONAL` {?`b` `receives` ?`award`}

and an RDF graph $G = \{(a, \texttt{writes}, b)\}$. Assuming that $models(G)$ contains all supersets of G (like e.g. under the RDF simple semantics), $certain(Q, G) = \emptyset$.

* This work was supported by the Austrian Science Fund (FWF): P30930-N35

Given that $(a, \texttt{writes}, b) \in G'$ for every $G' \in \mathit{models}(G)$, this is an unintuitive result. However, the mapping $\mu = \{(?\texttt{auth}, a)\}$ is no certain answer because it is not part of $Q(G')$ for every $G' \in \mathit{models}(G)$. For example, take $G' = G \cup \{(b, \texttt{receives}, p)\}$. Then $Q(G') = \{\{(?\texttt{auth}, a), (?\texttt{award}, p)\}\}$.

What makes this situation unintuitive is that while $\mu \notin Q(G')$, an *extension* of μ is contained in $Q(G')$. In fact, weakly monotone queries are exactly those queries Q where for all pairs of RDF graphs $G \subseteq G'$ the result $Q(G')$ contains a not necessarily proper extension of every mapping in $Q(G)$ [3].

To account for this, based on the notion of subsumption (a mapping μ' subsumes a mapping μ if μ' extends μ), an alternative certain answer semantics was proposed in [1]. One challenge in devising such a semantics is to avoid introducing unjustified subsumed mappings to the certain answers.

Example 2. Consider the query Q from Example 1 and let the RDF graph G_1 be the RDF graph G' from Example 1. Assuming $\mathit{models}(G_1)$ to contain all supersets of G_1, one would expect $\mathit{certain}(Q, G_1) = \{\{(?\texttt{auth}, a), (?\texttt{award}, p)\}\}$, while there is no justification for $\{(?\texttt{auth}, a)\} \in \mathit{certain}(Q, G_1)$. However, for $G_2 = G_1 \cup \{(a, \texttt{writes}, c)\}$, intuitively $\mathit{certain}(Q, G_2) = \{\{(?\texttt{auth}, a), (?\texttt{award}, p)\}, \{(?\texttt{auth}, a)\}\}$. This is because, unlike for G_1, over each graph in $\mathit{models}(G_2)$ at least two different mappings contribute a solution, namely $\{(?\texttt{auth}, a), (?\texttt{b}, b), (?\texttt{award}, p)\}$ and $\{(?\texttt{auth}, a), (?\texttt{b}, c)\}$. However, due to projection, these two mappings may lead to the same solution: for $G_3 = G_2 \cup \{(c, \texttt{receives}, p)\}$ we get (under set semantics) $Q(G_1) = Q(G_3)$. Under bag semantics, $Q(G_1)$ and $Q(G_3)$ still contain the same mapping, but it occurs once in $Q(G_1)$ and twice in $Q(G_3)$.

As a result, these two cases cannot be distinguished under set semantics, which prevents a definition of certain answers that acknowledges the differences between these cases. In [1] this was resolved by excluding all subsumed mappings from the certain answers. E.g., in the above example, $\{(?\texttt{auth}, a), (?\texttt{award}, p)\}$ would be the only certain answer for both, G_1 and G_2. While being a sensible definition, it is nevertheless a little ad hoc.

Given recent advances in SPARQL query answering and reasoning under bag semantics (cf. [6, 9, 2, 4, 5]), in this ongoing work we are revisiting the definition of a certain answer semantics for SPARQL under *bag semantics*, with the final goal to devise a certain answer semantics that (more) adequately describes the certain information returned by weakly monotone queries. This submission does not present new results, but suggests a possible certain answer semantics, (hopefully) showcasing that revisiting certain answer semantics for SPARQL is worthwhile.

2 Subsumption between Bags

A possible way of defining a certain answer semantics that faithfully includes subsumed mappings is to introduce a "subsumption-aware" variant of bag-intersection.

Towards this goal, we fix some notation. A mapping μ is a set of pairs $(?\texttt{x}_\texttt{i}, v_i)$, each pair denoting $\mu(?\texttt{x}_\texttt{i}) = v_i$. A bag M of mappings is a collection of mappings

that may contain each mapping more than once. We write $card_M(\mu)$ to denote the number of times a mapping μ occurs in bag M (if clear, M may be dropped; if we do not specify $card_M(\mu)$, we assume 1 by default). In this submission, we will assume RDF graphs to be sets, while we assume query results to be bags of mappings, a setting sometimes referred to as set-bag semantics in the literature.

Having settled this, we first have to extend the notion of subsumption to bags. For sets L, R of mappings, subsumption $L \sqsubseteq R$ holds when for every mapping $\mu \in L$ there exists a mapping $\mu' \in R$ such that $\mu \subseteq \mu'$. Similar to homomorphisms (cf. [7]), there are several possibilities for extending subsumption to bags L, R of mappings: one could just apply the definition for sets, or one could demand that for every mapping $\mu \in L$ there exists $\mu' \in R$ such that $\mu \subseteq \mu'$ and $card_L(\mu) \leq card_R(\mu')$. While it would be interesting to study the effects of these definitions, for our purpose they are too weak. For example, they cannot resolve the situation described in Example 2: under both defintions, $Q(G_1)$ and $Q(G_2)$ would subsume each other. Thus a subsumption based definition of certain answers could not distinguish $Q(G_1)$ from $Q(G_2)$, despite our intention that $certain(Q, G_2) = Q(G_2)$, but not $Q(G_1)$.

We thus use a stricter definition for subsumption between bags, and say that a bag L is subsumed by a bag R, written $L \sqsubseteq_b R$, if there exists a mapping $h\colon L \to R$ such that $\mu \sqsubseteq h(\mu)$ for all $\mu \in L$ and $card_R(\mu') \geq \sum_{\mu \in L\colon h(\mu)=\mu'} card_L(\mu)$ for all $\mu' \in R$ (this corresponds to the additive homomorphisms in [7]).

Next, we say that a bag M of mappings is $\sqsubseteq_b$ maximal w.r.t. a property Ω if M satisfies Ω and there is no M' satisfying Ω such that $M \sqsubseteq M'$ but $M' \not\sqsubseteq M$.

It is an interesting observation at this point that sets $M_1 \neq M_2$ of mappings may satisfy $M_1 \sqsubseteq M_2$ and $M_2 \sqsubseteq M_1$ (just consider the mappings in Example 2), while (for bags or sets) $B_1 \sqsubseteq_b B_2$ and $B_2 \sqsubseteq_b B_1$ implies $B_1 = B_2$.

The notion of $\sqsubseteq_b$-maximal bags now allows us to define $L \cap_\sqsubseteq R$, a version of bag-intersection that retrieves maximal information from both, L and R: for two bags L, R of mappings, let $L \cap_\sqsubseteq R$ be a $\sqsubseteq_b$-maximal bag M such that $M \sqsubseteq_b L$ and $M \sqsubseteq_b R$. Unfortunately, the result of this operator is not necessarily unique.

Example 3. Let $L = \{\{(x,1),(y,1),(u,1)\},\{(v,1)\}\}$ and $R = \{\{(x,1),(y,1),(v,1)\}, \{(u,1)\}\}$ be two bags of mappings. Then $M_1 = \{\{(x,1),(y,1)\}\}$ and $M_2 = \{\{(u,1)\},\{(v,1)\}\}$ are both $\sqsubseteq_b$-maximal w.r.t. being subsumed by L and R.

However, we will discuss next that in many cases $\cap_\sqsubseteq$, or a slight adaption of it, is nevertheless well-suited to define meaningful certain answers.

3 Certain Answers via $\cap_\sqsubseteq$

Besides not returning a unique bag, another property of $\cap_\sqsubseteq$ needs to be taken care of before it can be used to define certain answers, as illustrated next.

Example 4. Consider a SPARQL query Q

`SELECT ?a, ?w, ?r WHERE {?a isa author} OPTIONAL {?a writes ?w.?a reads ?r}`

and an RDF graph $G = \{(a, \texttt{isa}, \texttt{author}), (a, \texttt{reads}, b)\}$. Assume that also the

knowledge "every author writes some book" (expressed e.g. in OWL) is given and that every RDF graph $G' \in \mathit{models}(G)$ satisfies this condition. Then $\bigcap_{\sqsubseteq}\{q(G') \mid G' \in \mathit{models}(G)\} = \{\{(?\mathtt{a}, a), (?\mathtt{r}, b)\}\}$.

However, this result does not respect the requirement expressed in the query that a result should contain either a value for both, $?w$ and $?r$, or neither of them. As a result, instead of defining certain answers just as $\bigcap_{\sqsubseteq}\{q(G') \mid G' \in \mathit{models}(G)\}$, following [1], we also restrict the possible domains for the certain answers. For a set $\mathcal{V}$ of sets of variables, we therefore extend $L \cap_{\sqsubseteq} R$ to $L \cap_{\sqsubseteq}^{\mathcal{V}} R$ as being the $\sqsubseteq_b$-maximal bag M such that $M \sqsubseteq_b L$, $M \sqsubseteq_b R$, and $\mathit{dom}(\mu) \in \mathcal{V}$ for all $\mu \in M$.

Finally, for a query Q, let the *admissible solution domains* $\mathit{adsoldom}(Q)$ be the set of possible domains of mappings in $Q(G)$ (for any G). Due to space restrictions we stick to this vague definition; but, for example, for conjunctive queries Q, $\mathit{adsoldom}(Q)$ contains as single element the set of all output variables of Q, for query Q from Example 2 we get $\mathit{adsoldom}(Q) = \{\{?\mathtt{auth}\}, \{?\mathtt{auth}, ?\mathtt{award}\}\}$, and for Q from Example 4, $\mathit{adsoldom}(Q) = \{\{?\mathtt{a}\}, \{?\mathtt{a}, ?\mathtt{w}, ?\mathtt{r}\}\}$.

Definition 1. *Let G be an RDF graph, Q a query, and $\mathit{models}(G')$ the set of graphs entailed by G. Then the* certain answers *of Q are defined as*

$$\mathit{certain}(Q, G) = \bigcap_{\sqsubseteq}^{\mathit{adsoldom}(Q)}\{Q(G') \mid G' \in \mathit{models}(G)\}.$$

While, for arbitrary inputs $L, R, \mathcal{V}$, the result of $L \cap_{\sqsubseteq}^{\mathcal{V}} R$ is not necessarily unique, in most of the settings in which we compute certain answers, L, R, and $\mathcal{V}$ are not arbitrary inputs but adhere to some structure that we can exploit.

For example, when applied to conjunctive queries, $\cap_{\sqsubseteq}^{\mathcal{V}}$ reduces to conventional (set- or bag) intersection, and as a result for these queries we get the "classical" definition of certain answers as a special case of Definition 1.

Similarly, in the special case of $\mathit{models}(G)$ containing a minimal element G (i.e. $G \subseteq G'$ for all $G' \in \mathit{models}(G)$), for weakly monotone SPARQL queries the certain answers are uniquely defined. In fact, applied to the settings in Examples 1 and 2 it produces exactly the intuitive bags of certain answers.

More generally, also for the sets $\mathit{models}(G)$ that exhibit a *canonical model* (i.e. that contain some $G' \in \mathit{models}(G)$ such that for every $G_i \in \mathit{models}(G)$ there exists some homomorphism $h_i \colon G' \to G_i$) we strongly conjecture that for weakly monotone queries, and especially for well-designed SPARQL queries, Definition 1 gives a unique bag of certain answers.

There are, of course, a lot of open question for future and ongoing work. These include the relationship to certain answer semantics from the literature (e.g., while "typical" certain answers for CQs are a special case of Definition 1, this seems not to be the case for the definition in [1]), an investigation of classes that provide a unique bag of certain answers, and of course the costs/complexity of query evaluation under this semantics for specific reasoning formalisms (e.g. OWL). Another line of research is to establish a connection with the work in [8]. Finally, also considering alterations of this definition could be of interest.

References

1. Ahmetaj, S., Fischl, W., Pichler, R., Simkus, M., Skritek, S.: Towards reconciling SPARQL and certain answers. In: Proc. WWW 2015. pp. 23–33. ACM (2015)
2. Angles, R., Gutiérrez, C.: The multiset semantics of SPARQL patterns. In: Proc. ISWC 2016. LNCS, vol. 9981, pp. 20–36. Springer (2016)
3. Arenas, M., Pérez, J.: Querying semantic web data with SPARQL. In: Proc. PODS 2011. pp. 305–316. ACM (2011)
4. Console, M., Guagliardo, P., Libkin, L.: Approximations and refinements of certain answers via many-valued logics. In: Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016. pp. 349–358. AAAI Press (2016)
5. Console, M., Guagliardo, P., Libkin, L.: On querying incomplete information in databases under bag semantics. In: Proc. IJCAI 2017. pp. 993–999. ijcai.org (2017)
6. Hernández, D., Gutierrez, C., Hogan, A.: Certain answers for SPARQL with blank nodes. In: Proc. ISWC 2018. LNCS, vol. 11136, pp. 337–353. Springer (2018)
7. Hernich, A., Kolaitis, P.G.: Foundations of information integration under bag semantics. In: Proc. LICS 2017. pp. 1–12. IEEE Computer Society (2017)
8. Libkin, L.: Certain answers as objects and knowledge. vol. 232, pp. 1–19 (2016)
9. Nikolaou, C., Kostylev, E.V., Konstantinidis, G., Kaminski, M., Grau, B.C., Horrocks, I.: The bag semantics of ontology-based data access. In: Proc. IJCAI 2017. pp. 1224–1230. ijcai.org (2017)
10. Pérez, J., Arenas, M., Gutiérrez, C.: Semantics and complexity of SPARQL. ACM Trans. Database Syst. **34**(3), 16:1–16:45 (2009)