Answer to the reviewers

We wanted to thank the reviewers for their useful comments that allowed us to improve our paper. Following our detailed answers for each of their points. We used two colors, both here in this answer and in the paper, to identify the answer to each comment.

Reviewer 1

The font of the answers to reviewer 1 are set with this color both in the rebuttal letter and in the paper to highlight modifications.

Summary

The paper is about data credit that represents the importance of data cited by a paper or another research entity. The authors introduce new distribution strategies (DS) that are based on lineage and why and how provenance. They evaluate these strategies over real-world and synthetic queries over GtoPdb found in papers published in the British Journal of Pharmacology (BJP). Furthermore, the DSs are compared against traditional citations to prove that the credit appears to be a useful new measure.

Strong points

S1) The paper introduces a new method for valuing the impact and importance of data in the research world with clear motivation.

S2) The authors provide extensive experiments over both real-world and synthetic queries.

Thank you.

Weak points

W1) The technical contribution may not be strong enough. The credit distribution in a sense follows the property of the provenance types.

We agree that the credit distribution strongly depends on the provenance being used to define it.

Still, one of the paper's main contributions consists of defining the concept of data credit and alternative methods for distributing it by using diverse forms of provenance. We provide a viable system for calculating and distributing data credit and we show how this differs from the state-of-the-art in informetrics which is based on citation count.

Overall, we also present a new application for data provenance.

In this updated version we also added two new distribution strategies, as suggested by Reviewer 2, inspired by the concept of Responsibility and the Shapley value. Even though related to data provenance, these two strategies are a completely new form of credit distribution. As such, we integrated the experiments and the observations with this new DS.

W2) Some settings for experiments are not clear (see the detailed comments below).
Thanks for the observation. We faced each point of the detailed comments below.

W3) Writing can be improved significantly, e.g., it is quite redundant especially on the explanations of data provenance.

As above, we tackled the single comments on the writing.

Detailed comments
D1) In the introduction, the motivation is clear but it is a bit long dragging to the summary of the contributions.

We worked on the introduction, reducing the explanation of the data provenances as suggested and streamlining the discussion. We also removed some paragraphs reporting background information about data citations that are better described in the Background section.

Furthermore, it is not clearly introduced that how the new DSs are developed based on why and how provenance.

We described in the introduction that we use these forms of provenance (why- and how-provenance and now also responsibility and Shapley value) to define new Distribution Strategies that have a different behavior with respect to the one based on lineage. The newly defined DS strategies exploit the additional information provided by these forms of provenance to weight the contribution of the tuples on the basis of their role in producing the output.
We then discuss why one DS may be preferred to another depending on the application and its goals.

D2) In Section 2, what is the difference between the credit computation process and distribution process?

We added the following paragraph in Section 2 about the difference between the two processes:
“Therefore, when discussing data credit, we need to consider credit computation -- i.e., the process to compute the quantity of credit generated by the citation -- and credit distribution -- i.e., the process to distribute credit and to assign it to the entities that contributed to the creation/curation of the cited data. In this paper we focus on the latter.”

D3) The different types of provenance are mentioned in multiple places, e.g., introduction, Section 2.3, and Section 4 again. Even

Section 5 also mentions the different types of provenance as explaining the new DSs.

We reduced the redundancy in the use of the notions of provenances through the paper as indicated by the reviewer.
In particular, we reduced the content of section 2.3, moving the technical aspects to section 4. Finally, in Section 5 we removed the notions about provenance that we already discussed in Section 4.

D4) Section 5 introduces technical contributions by providing several definitions of each DS.

Yes.

D5) In Section 5.4, the authors explain how-provenance based DS and introduce several notations, e.g., c(H), c(M_2), mc, and e, which make the reader confusing such that c(H) means coefficient and c(M_2) represents exponents of the monomial. I wonder why c and e can't be used for coefficient and exponent, respectively.

We agree with the reviewer that the notation may be confusing. We changed the name of the second function *c* to *e*. We also changed the name of the previous function *e* to *te*. We added a table (Table 5) with the notation used in Section 5, which we hope helps to better follow this part.

D6) On page 24, line 648, the credit for f_1 is 1/2 and 1/4 for others. Isn't it 1/4 for f_1 and 1/8 for others?

The reviewer is correct, but we were talking in terms of proportions of credit, not absolute values of credit being assigned. Thus, assuming that the monomial is f^2*c2f * c, ½ of the credit given to this monomial is given to f (thus ½ * ½ = ¼, as correctly noted by the reviewer), and ¼ is given to the other two tuples. We removed this sentence since it was not essential for the understanding of the approach and to avoid confusion for the readers.

D7) In the experiment section, the credit for an output tuple is assigned to 1. How does the evaluation change if the output tuple is assigned with unequal credit? Why does the model need to be modified (assume that since it is clarified as future work)?

We assume that with “unequal credit”, the reviewer is asking what changes when tuples in the output may be carrying credit different than 1.
If this is the case, the model does not need to change, i.e., the formulas for credit distribution stay the same, since they are defined for a generic value of credit k. Of course, if the credit to be distributed in input changes, the output distribution will also change, resulting in some tuples of the database receiving more credit than others. There is no need, therefore, to change the model itself. Of course, we are focusing here on the distribution model, we did not talk in our paper about the credit generation criteria.

We added a new section, Discussion, where we discuss this aspect in a paragraph called “Credit Generation”.

D8) How are the 10K synthetic provenance polynomials generated randomly? Similarly, how are the ten tuples are selected in Figure 10?

When we wrote “randomly”, we meant picking queries from a uniform distribution of synthetically generated provenance polynomials; we added a further explanation in the paper. The polynomials were generated as follows : we first decide how many monomials will constitute the polynomial, a number between one and six. A monomial with coefficient 2 will count as 2 monomials. We then randomly take a tuple from the table family, one from the table contributor2family and one from table contributor, acting as variables for the monomial. We randomly choose a coefficient for this monomial (between 1 and 3) and an exponent for each tuple (between 1 and 4). For the next monomial, then, we decide if we want to keep the same tuple from the table family as the first tuple of the new monomial. To do so, we generate a random floating number between 0 and 1. If the value is above 0.2, we change the family tuple.

We added the description above to the paper.

The tuples were chosen among all the possible tuples to show the evolution of credit in time, so we observed the heat maps corresponding to the different moments and we cherry-picked ten tuples to get a meaningful sample to show what may happen by using different DSs.

D9) As shown in Section 6.1, most of the queries used in the real world are quite simple. In this case, lineage-based and why and how provenance-based DSs return the same credit distribution. Then, what would be the best choice, lineage or why provenance based? Furthermore, I wonder how often complex queries are used in the real world.

We added a discussion section where we also discussed the possible choices for a DS.
In a few words: lineage may be used by users only interested in identifying the presence of relevant parts of the database. Why-provenance can be used instead when it is required that the parts of the database that are more critical for the generation of the output are rewarded more. This may be the case for applications that use credit as support for other steps. More details about the other three DSs were also added in the paper.

It is also worth mentioning that, while in our paper we first computed the how-provenance and then derived the other provenances from that one, in a general case scenario users may prefer lineage because it is easier to implement with respect to why-provenance and how-provenance. That may be another reason why lineage may be preferred.

Moreover, as noted by the reviewer, in certain cases the two DSs may behave the same due to the nature of queries, thus making the choice de-facto irrelevant . This is, however, not always the case.
Although we did not find any work that shows the type of provenances found in real-world queries, there are other papers in the literature that analyze the nature of query logs and show a certain amount of complexity in the nature of queries. In particular, why-provenance and how-provenance, as distribution strategies, potentially differ from the distribution obtained from lineage when joins and projections operations are included in the queries.

We cited these works and commented on them in the Discussion section.

These results make us think that complex queries with joins and projections are actually adopted by users, and thus different types of distributions can actually be performed.

D10) Minor:
- page 5 line 110: CDC -> DCD
- p6 l145: give -> gives
- p9 l229: WG is used before defined

We fixed these typos, thanks.

- Coloring issue in printed version, e.g., in p31 at l806, the light blue is not distinguishable.

We changed the color to yellow, which should now be distinguishable when printed.

- p35 l907-909: if the tuples from three tables are randomly chosen, wouldn't the more collaborative authors be selected?

Actually this is a synthetic experiment. We are randomly assigning new authors to tuples. Thus, possibly, we are assigning to the same synthetic author tuples that, in reality, are authored by different groups of authors.

Reviewer 2

The answers to this reviewer were made here and in the paper with this color.

Reviewer 2: The authors present an approach for distributing credit from a query result tuple to the inputs of the query. This problem arises when the credit for a citation to a dataset should be fairly distributed to contributors that produced part of that dataset. The authors propose to use provenance as to determine which input should receive credit and how much. Three distribution strategies based on three provenance types (lineage, why-provenance, and provenance polynomials) are presented in this paper out of which one was already presented by the authors in [24]. While this is sensible, there are some issues with the technical development (see below). The authors experimentally evaluate their approach over a real world curated database. However, the evaluation mainly focuses on comparing the credit distributions created by the three strategies.

There is not much that can be learned from that. I think comparing against some type of ground truth (e.g., collected by asking authors to rate the importance of parts on their research) would be more meaningful.

A ground truth defining the best possible credit distribution is something very much desirable but impossible to obtain. Indeed, there is no perfect credit distribution that works for all the involved stakeholders. As a comparison, let us think about citation distribution. In bibliometrics, the transitive closure of citations is a long studied problem with no unique solution; if I cite paper A which in turn was based on paper B and paper C, how many citations (or credit portions) should paper B and paper C receive for their contribution? In the current system, the answer is no citations and one reason is the difficulties to find an agreement about the best distribution of credit.
In the realm of data, this problem is enhanced. In fact, it is hardly possible that a scientist who contributed to some data in a scientific database has an idea about how a query to the DB is answered, the data involved and how to weigh the relevance of the involved data.
If we ask paper authors to rate the data they used, they might provide some answers about the processed data they consumed, but they hardly have an idea about the data involved in the process that probably they have never seen.

Also, we note that asking scientists to understand how a query works, what are the data involved and how they are transformed is a daunting task with limited possibilities of success. Data provenance techniques have exactly this purpose: to track and materialize what happens under the hood and to make it evident to data experts.

Moreover, If we asked an author to rate their work, they could choose some data that they deem important because of their own liking or bias. The process of data credit distribution, on the other hand, enables us to find that, possibly, data that were not kept in high regard by their authors might instead have a big value in the research community because of dynamics that are hidden from human observation (and vice-versa). This possibility of discovering the impact of (possibly hidden) data, independently from the opinion that the authors have of their own work and from any other criteria if not the presence of data citations, is, in our opinion, one of the main potentialities of credit distribution.

```
Furthermore, the way it is presented, it seems that using the most
detailed distribution strategy (why-provenance, because provenance
polynomials only are sensible under bag semantics) would always be
preferable unless there is some other disadvantage that was not
discussed.
```

We agree with the reviewer that polynomials are sensible only with bag semantics, but that is the semantics in which we are working. In our paper we focus primarily on conjunctive queries and, in particular, on queries performed in SQL. In general, we need to assume SQL to work in the bag semantics since the keyword DISTINCT is not always used.
For this reason, the information provided by how-provenance can actually be different and richer than the one provided by why-provenance.

About the preference of one provenance against the other, we added a paragraph in section 7 discussing the reasons why someone may prefer one called “Choice of Distribution Strategy”. Lineage is the simplest form of provenance (also in terms of implementation), thus it is possible that, if a user is only interested in tracking the areas of the database that are used by queries, it may be enough for their needs.
In case of applications that are more sensible to the quantity of credit being given to data, such as data pricing, why-provenance and how-provenance can be useful. They are also harder to implement, since they need to keep track respectively of witnesses and polynomials.

As we describe in the next points, we also implemented two DSs strategies based on responsibility and Shapley value as suggested by the reviewer. Responsibility, being based on causality, may be harder to compute since causality may be NP-hard to find (more details are reported in the paper). Moreover, as we showed in the paper, a DS based on responsibility acts similarly to the one based on why-provenance, so in this case there may not be good reasons to choose it.
For the case of Shapley value, very recently, an efficient implementation for Boolean queries (queries that output true or false, or 1 or 0) has been provided in Deutch, D., Frost, N., Kimelfeld, B., and Monet, M. (2022). “Computing the Shapley Value of Facts in Query Answering”, arXiv, 2112.08874, both in terms of an exact computations (which in practice works well for most queries) and in inexact one (which is extremely fast and provides the same ranking of tuples as the exact computation, but not necessarily the same values).

Overall, I really wanted to like this paper, because it tackles an interesting, novel problem. However, the technical development and experiments are disappointing. Nonetheless, I believe giving the authors the chance to improve their work in a major revision is sensible in my opinion.

Strong points
- Interesting new application of provenance
- The authors use a real world dataset and queries (in some experiments) which is great

Thank you, we appreciate your interest, the time you dedicated to the paper and the feedback that you gave us.

Weak points
- The improvement over [24] just barely meets the bar for a journal extension of a previous paper

We improved our paper with now two new forms of distribution strategy, based on Responsibility and Shapley value as suggested. See the next points for more details about this aspect.

The two new distribution strategies are problematic:
Why-provenance: It is unclear why why-provenance instead of minimal why-provenance is used, because it was shown that why-provenance may have irrelevant witnesses. This was further solidified in the work on K-relations which demonstrated that PosBool(X) which is equivalent to minimal why-provenance has the same equivalences as set semantics and, thus, is the right choice of provenance for set semantics.

We are not sure about what the Reviewer refers to when they talk about “minimal why-provenance”. We think they are referring to “minimal witness basis”, as it is described in Cheney J. et al, “Provenance in Databases: Why, How and Where”, in pages 6 to 10.

If this is the case, we note that this minimal witness basis is a particular basis made by witnesses that are minimal. In particular, the minimal witness basis is contained in all the witness basis of equivalent queries.

However, why-provenance is different, since it is a particular witness basis that, as described in the same paper, depends on the nature of the query. Therefore, two equivalent queries may have two different why-provenances due to the fact that the operations that they perform on the data are different, resulting in a different result of the algorithm computing the why-provenance. In other words, while the minimal witness basis depends on the semantics of the query, why-provenance depends on its structure.

We decided to use why-provenance exactly for this quality of depending on the operations actually performed by the query. The goal, in fact, is to distribute credit to the data actually being used by the query, even if there is some other equivalent query with a smaller witness basis.

As we notice in other points, our work is defined in the set of polynomials N[X], where the variables are taken from the set X and the exponents and coefficients from the set of natural numbers N, and not in PosBool(X).
We agree with the Reviewer that in PosBool(X) the minimal witness basis may be the more sensible choice as a form of provenance.

How-provenance: Provenance polynomials is the right choice for bag semantics, but not for set semantics, because additional equivalences hold for set semantics that do not hold for provenance polynomials (and bag semantics). For instance, under PosBool(X) (set semantics) the following equivalence holds: a + a * b = a. So in this case b is irrelevant for producing the output and all credit should be given to a. Either drop the provenance polynomial strategy or only apply it for bag semantics.

We agree with the Reviewer that in the PosBool(X) semiring, how-provenance is not a sensible choice.
As specified in the previous point however, we are working in a rather different context: with SQL queries, and thus bag semantics. This is true throughout the paper. In particular, the equivalence a + a*b = a holds in PosBool(X) since it is a semiring of absorptive polynomials, which is not true in general with other semirings. While it is true, therefore, that b in that case is irrelevant, that is not true in the case considered in the paper. The tuple b represents an alternative to the production of the output tuple, and as such we reckon that it deserves a portion of credit. While it is true that it is possible to define a distribution strategy that exploits the properties of a set such as PosBool, it is not in the scope of the paper to explore such possibility, since our aim was to exploit forms of provenance that were already well know and consolidated in the realm of relational databases and SQL queries rather than interesting but less known forms of provenance in other semirings.

The experimental evaluation mostly compares the credit distribution created by the different strategies. I do not think that this type of experiments leads to interesting insights apart from the obvious (more detailed provenance models better model tuple contribution). This becomes clear in even through provenance polynomials is not the right choice for these set semantics queries, this issue is not clear from the results. I think it would be more interesting to create some type of ground truth (maybe using human raters for influence if there is no other way to create a meaningful ground truth) and compare against that. Furthermore, there should be a discussion on why one would choose the lineage strategy at all (e.g., performance).

We note in a previous point how it is not possible, in the type of context considered in the paper, to really define a form of ground truth. As pointed out in a previous point for Reviewer 1, we added a paragraph in Section 7 on the pros and cons of the different provenances.

-The purpose of a credit distribution strategy is to determine the amount of influence an input tuple has on the existence of a query result. This problem has been studied in work on causality under the name of responsibility [1] and is also the motivation for Shapley values for database query answers [2]. The proposed approach should be compared against these metrics and it should be evaluated what the advantages and disadvantages of the proposed technique are.

[1] The Complexity of Causality and Responsibility for Query Answers and Non-Answers. A. Meliou, W. Gatterbauer, K.F. Moore, D. Suciu. Proceedings of the VLDB Endowment, 2010.
[2] The Shapley Value of Tuples in Query Answering. Ester Livshits, Leopoldo E. Bertossi, Benny Kimelfeld, Moshe Sebag. 23rd International Conference on Database Theory, ICDT 2020, March 30-April 2, 2020, Copenhagen, Denmark (2020).

We thank the Reviewer for providing us the pointers to the papers describing causality, relevance, and the Shapley Value.

We implemented two new distribution strategies based on responsibility and Shapley value and we described them in our paper together with the other three strategies. We provided the definition of the new responsibility-based DSs and we repeated the experiments to also include this last DS.
We re-created the synthetic polynomials and re-built the experimental section taking into consideration the new strategy.
As can be seen in more detail from the discussion in the paper, responsibility appears to behave in a way very similar to why-provenance. Thus, being in the general case hard to compute, it may not be the best decision for a DS.
The Shapley value instead behaves differently, bringing more diversity among the proposed DSs.

In the paper we added section 4.4, that introduces causality and responsibility, section 4.5, that introduces the Shapley value for facts in a relational database, and sections 5.5 and 5.6 that describes the new DSs. We also performed again all the experiments based on synthetic polynomials, distributing credit using also this last DS, and included the results and the comments in the paper.

While it is reasonable that this paper focuses on the subproblem of distributing a given amount of credits for a query answer to the inputs of the query, it would have been good to outline what problems remain to be solved for the distribution strategies to be employed to distribute citation credits. For instance, what are possible methods for deciding how much credit to assign to a query answer and what are their issues (e.g., authors assigning too much credits to self-citations)?

In the first two paragraphs of the section Discussion we talked about the problems of generating credit , how to decide how much credit to generate and how to propagate it. We preferred to leave out the discussion about self citations since it is not really a data-specific problem.

Detailed comments

-I think the introduction of the provenance models should be more formal

While we agree with the reviewer about the necessity for formality, we also reckon that this paper is not theory-focus, but more grounded on experiments and potentiality of credit distribution. The paper has more than 50 pages, with many definitions, formulas and experiments.
We thought that adding more formality to the introduction of the models, while the formal definitions are already present in the corresponding sections and related papers, would inficiate the overall readability of the paper.

- In the discussion of the provenance types, the authors state that provenance polynomials have the advantage that they track how input tuples are combined to produce the answer and that the same is not true for why-provenance. However, as mentioned before why-provenance is equivalent to PosBool(X) which tracks this information. In fact, the set of witnesses is a disjunction of options (addition in semirings) and each witness is a conjunction (all tuples from the witness together produce the results which is multiplication in semirings).

We addressed this observation in a previous point. We think, therefore, that the use of how-provenance remains sensible. This is an observation that is also supported by the results of our experiments with synthetic polynomials, that show how the distribution based on how-provenance actually brings different results from the one with why-provenance. This is also true with all the other tested DSs: they all behave differently on the synthetic provenance polynomials.