SIGMOD2015
**SIGMOD2015**
May 31- June 4, 2015, Melbourne, Australia

**Reviews For Paper**

| | |
|---|---|
| **Track** | Research 2nd Submission (Revision) |
| **Paper ID** | 293 |
| **Title** | Identifying the Extent of Completeness of Query Answers over Partially Complete Databases |

**Masked Reviewer ID:** Assigned_Reviewer_1

**Review:**

| Question | |
|---|---|
| Rating for the *original* submission | Revise and Resubmit |
| Summary of the paper (prior to revision) | The authors propose a form of metadata describing the completeness of a relational dataset. An algebra over this metadata is constructed in parallel to SPJU relational algebra, which allows users to obtain the completeness of query results. The authors prove completeness and soundness of the algebra, and show the results of experiments on its runtime characteristics.<br><br>The proposed approach is quite nifty, but overlaps substantially with existing work (that is not referenced). Moreover, while most of the paper is written quite well, there are portions that are extremely difficult to parse. |
| Three (or more!) strong points about the *original* version of the paper | - Great idea, exploring a very important space<br>- Rigorous formal treatment of the problem<br>- Elegant solution |
| Three (or more!) weaknesses of the *original* version of the paper | - Does not differentiate itself sufficiently from prior work<br>- Focuses on a very specific sub-problem... see below.<br>- Section 5 is written very sloppily, and overcomplicates a relatively simple problem. |
| Is the paper relevant for SIGMOD? | Yes |
| Significance | The paper improves on existing work |
| Technical depth and quality of content | Solid work |
| Validation - experiments and proofs | Very nicely support the claims made in the paper |
| Presentation | Reasonable: improvements needed |
| Discussion of related work - recall that the new page limits | |

| | |
|---|---|
| allow for material outside the 12pp including references, hence we expect good coverage of related work | Inadequate description of related work |
| Detailed evaluation of the *original* version of the paper | My big issue with the paper is that it occupies a space almost identical to the SIGMOD 2014 paper, "Partial Results in Database Systems" by Lang et al. The authors of the submitted paper explore a specific case of Lang et al.'s solution in much greater depth, distinguishing them somewhat. However, given the level<br>of overlap between the two and the fact that at a glance, Lang's solution seems to be a more general form of the submitted paper, I would expect to see a<br>MUCH more detailed discussion of how the two differ.<br><br>My other major issue is Section 5, which is full of typos, terms that are not properly defined, solutions to problems that are not characterized, and outright<br>sloppy writing that makes it virtually impossible to follow what the section is trying to do or why.<br><br>----------<br>Nitpicks<br><br>- The term 'Punctuations' is used without being defined.<br>- The kerning in the SELECT query before equation 1 is ugly. Use a fixed width font.<br>- The discussion of incomplete databases in the related work section could be expanded a bit -- Certain and possible answers do form a large portion of the area, but identifying and quantifying possible errors is a large part of it as well.<br>- You assume categorical data. This is an extremely strong assumption that drives many of your design decisions, and breaks down as you begin to generalize your approach (e.g., if you include Aggregates)<br>- At the start of 4.1, make it easier for your readers. Append a "(e.g., \tilde{\sigma})" to the sentence that reads "We make the distinction ... by adding a tilde character to the latter."<br>- S 4.1.1 has a typo: "For the selection \sigma_{A=B}" should be \sigma_{A=d}<br>- S 4.1.3 for consistency: "For a selection A=B" should be \sigma_{A=B}<br>- S 4.1.4 you might want to define P_{maint} and P_{teams} instead of the awkward over-line.<br>- The paragraph right before Example 8 ("We do not want to compute patterns...")<br>is a very awkward transition. Are you missing a subsection header?<br>- Same deal for the paragraph right before S 4.2. It seems like 4.2 and 4.3 should be part of an experiments section.<br>- S 4.2: "We also assume that... attributes with a relatively low number of distinct values are used". This seems like a pretty wild assumption, made mainly because it works well with your system. Even the examples that you gave do not really follow this property.<br>- Table 5 has multi-line subscripts for the join and select conditions. This formatting is extremely hard to parse.<br>- The font size in Figures 4 and 5 is almost unreadable. |

| | |
|---|---|
| | ----------<br>Section 5<br><br>- Sentence 1: "saSqw"?<br>- 5.1 does a bad job of motivating each of the individual building blocks. It's hard to figure out what 'unify', 'promote', etc... do if you don't tell me what the high-level goal is. I hate it when people tell me this, but the section really needs to be written more top down. If A and B form the entire allowable domain for the 2nd attribute, then<br>{ (*, A, *, *, *), (*, B, *, *, *)} = (*, *, *, *, *)<br>Great, so how do I find out what the allowable domain of an attribute is? I'm on-board, excited, and ready to find out the answer... but instead what I get is a slew of new terms and some functions without any intuition as to why those terms and functions will help me understand your solution.<br>- In general, your solution here seems to not only be presented in a convoluted<br>way, but also seems to itself be rather convoluted. Finding the allowable domain of an attribute is both more general and easier to understand than the somewhat more specialized problem of coalescing patterns. |
| What specific revisions did you seek from the authors? | (1) Differentiate yourselves from Lang et al, and (2) Completely rewrite Section 5. |
| Did the authors fully implement the changes requested for the revision? | Yes, I am satisfied with the changes made |
| Please provide detailed comments on the revision | All of the changes I was looking to see have been implemented to my satisfaction. Only a few minor nitpicks remain:<br><br>- Section 1, paragraph 5: "Despite these differences, in common between all three scenarios..." -- Bad grammar. Please fix.<br>- Section 4.2: "Considering average, this number is even much better at 2.4\%..." -- This sentence is confusing (as it itself admits) and does not any useful information.<br>- Section 4.2: A summary of the queries, and a table or graph showing detailed per-query results would be nice, even if it had to be deferred to the appendix. |
| Final overall rating, after revision | Accept |

**Masked Reviewer ID:** Assigned_Reviewer_2

**Review:**

| Question | |
|---|---|
| Rating for the *original* submission | Reject |
| | The paper addresses the problem of incomplete data in databases. Specifically, it proposes and assigns annotations on the completeness of the |

| | |
|---|---|
| Summary of the paper (prior to revision) | data in relational tables and shows how these annotations can be propagated to query results. The annotations and associated algebra are quite elegant theoretically. However, it is not clear what the practical applications are and hence it is not clear how to evaluate the proposed framework. |
| Three (or more!) strong points about the *original* version of the paper | 1. Addresses an interesting problem of characterizing the extent of completeness of relational data and the query answers.<br><br>2. Using the formalism of an algebra for the transformation of the annotations via relational algebra queries is quite elegant mathematically. |
| Three (or more!) weaknesses of the *original* version of the paper | 1. Not clear how the proposed annotations and algebra can be used in practice.<br><br>2. Not clear how to evaluate the proposed framework.<br><br>3. Experimental evaluation is consequently weak. |
| Is the paper relevant for SIGMOD? | Yes |
| Significance | The paper improves on existing work |
| Technical depth and quality of content | Syntactically complete but with limited contribution |
| Validation - experiments and proofs | Unclear/obscure, hard to determine what is going on and what has been validated |
| Presentation | Reasonable: improvements needed |
| Discussion of related work - recall that the new page limits allow for material outside the 12pp including references, hence we expect good coverage of related work | Clear explanation of the state of the art and how this paper relates |
| Detailed evaluation of the *original* version of the paper | 1. The proposed annotations and algebra is quite elegant mathematically, but the reader feels like it is a hammer looking for a nail. While it is interesting theoretically, it is not clear how this can be applied in practice or what problem it solves in a real system.<br><br>2. Consequently it is difficult to evaluate experimentally how effective the proposed framework is. The experimental results in the paper are good in that it explores and validates properties of the framework (minimization of completeness patterns etc), but the reader still does not know what the proposed framework is good for.<br><br>3. Perhaps it would really help the paper to have a motivating problem from a realistic application and to design additional experiments that links to the efficacy of the proposed framework for addressing a real problem. |

| | |
|---|---|
| What specific revisions did you seek from the authors? | 1. It would really help the paper to have a motivating problem from a realistic application<br>2. Design additional experiments that links to the efficacy of the proposed framework for addressing a real problem. |
| Did the authors fully implement the changes requested for the revision? | Partly, but not to my complete satisfaction |
| Please provide detailed comments on the revision | The revised version is much improved with the added motivating examples. While it is still not completely clear how useful the proposed method would be in a realistic setting, at least, the reader is given a glimpse. |
| Final overall rating, after revision | Neutral |

**Masked Reviewer ID:** Assigned_Reviewer_3

**Review:**

| Question | |
|---|---|
| Rating for the *original* submission | Accept |
| Summary of the paper (prior to revision) | This paper proposes a means of providing providing assertions about the completeness of a dataset wrt future updates in the form of so-called completeness patterns. Within this basic model the technical contributions of the paper include an algebra for computing completeness patterns for query results, refinements of the algebra to take the database instance into account to produce tighter completeness guarantees, and an experimental investigation into the practicality of the techniques. Compared to other papers in this area, the model seems perhaps to represent a better compromise between expressive power of the completeness model, and algorithmic tractability of the associated reasoning tasks. |
| Three (or more!) strong points about the *original* version of the paper | S1. Seems like a useful capability to offer for certain application domains.<br><br>S2. Rigorous and readable theoretical development.<br><br>S3. Model seems to strike a good balance between expressiveness and tractability.<br><br>S4. Nice use of techniques from outside the usual database systems toolbox (data structures from theorem proving) in the implementation. |
| Three (or more!) weaknesses of the *original* version of the paper | W1. Experimental sections are not as strong as the theoretical developments.<br><br>W2. Practical impact of the work is unclear. The problem feels a bit niche, and the solution gives a capability that seems nice to have but probably not crucial even within that niche. |
| Is the paper relevant for | Yes |

| | |
|---|---|
| SIGMOD? | |
| Significance | The paper improves on existing work |
| Technical depth and quality of content | Solid work |
| Validation - experiments and proofs | OK, but do not cover all of the claims |
| Presentation | Excellent: careful, logical, elegant, understandable |
| Discussion of related work - recall that the new page limits allow for material outside the 12pp including references, hence we expect good coverage of related work | Clear explanation of the state of the art and how this paper relates |
| Detailed evaluation of the *original* version of the paper | W1. A broader range of experimental workloads, involving real data, would strengthen the paper. The network dataset seems nice but the introduction of completeness annotations seems to have been done somewhat arbitrarily. Can this be done in such a way that connects with the workings of the application domain better? The TPC-H workload feels artificial.<br><br>There do not seem to be any running time numbers for the zombie patterns to try to quantify the price paid for the richer functionality.<br><br>W2. It's clear that the techniques of the paper yield a new capability absent from current systems, but it would make for a stronger story to explain in more concrete terms the ramifications of this new capability. What does being able to reason about the completeness of query results really buy in envisioned applications? |
| What specific revisions did you seek from the authors? | None requested. |