Causality in Databases, Database Repairs, and Consistency-Based Diagnosis

(extended abstract)

Leopoldo Bertossi and **Babak Salimi**
Carleton University, School of Computer Science
Ottawa, Canada. {bertossi, bsalimi}@scs.carleton.ca

When querying a database, a user may not always obtain the expected results, and the system could provide some explanations. Explanations that could be useful to further understand the data or check if the query is the intended one. Actually, the notion of explanation for a query result was introduced in [19], on the basis of the deeper concept of *actual causation*.

Intuitively, a tuple t is an *actual cause* for an answer $\bar{a}$ to a conjunctive query $\mathcal{Q}$ from a relational database instance D if there is a "contingent" set of tuples Γ, such that, after removing Γ from D, removing/inserting t from/into D causes $\bar{a}$ to switch from being an answer to being a non-answer. Actual causes and contingent tuples are restricted to be among a pre-specified set of *endogenous tuples*, which are admissible, possible candidates for causes, as opposed to *exogenous tuples*. (For a formalization of non-causality-based explanations for query answers in DL ontologies, see [3].)

Some causes may be stronger than others. In order to capture this observation, [19] also introduces and investigates a quantitative metric, called *responsibility*, which reflects the relative degree of causality of a tuple for a query result. In applications involving large data sets, it is crucial to rank potential causes by their responsibility [20, 19].

Actual causation, as used in [19], can be traced back to [11, 12], which provides a model-based account of causation on the basis of the *counterfactual dependence*. Responsibility was also introduced in [8], to capture the *degree of causation*. Apart from the explicit use of causality, research on explanations for query results has focused mainly, and rather implicitly, on provenance [4, 5, 7, 9, 15, 14, 23], and more recently, on provenance for non-answers [6, 13]. A close connection between causality and provenance has been established [19]. However, causality is a more refined notion that identifies causes for query results on the basis of user-defined criteria, and ranks causes according to their responsibility [20].

Consistency-based diagnosis [21] is an area of knowledge representation. The main task here is, given the *specification* of a system in some logical formalism and a usually unexpected *observation* about the system, to obtain *explanations* for the observation, in the form of a diagnosis for the unintended behavior.

In a different direction, a database instance, D, that is expected to satisfy certain integrity constraints (ICs) may fail to do so. In this case, a *repair* of D is a database D' that does satisfy the ICs and *minimally departs* from D. Different forms of minimality can be applied and investigated. A *consistent answer* to a query from D and wrt. the ICs is a query answer that is obtained from all possible repairs, i.e. is invariant or certain under the class of repairs. These notions were introduced in [1] (see [2] for a recent survey). We should mention that, although not in the framework of database repairs, model-based diagnosis techniques have been applied to restoring consistency of a database wrt. a set of ICs [10]

These three forms of reasoning, namely inferring causality in databases, consistency-based diagnosis, and consistent query answers (and repairs) are all *non-monotonic*. For example, a (most responsible) cause for a query result may not be such anymore after the database is updated. In this work we establish natural, precise, useful, and deeper connections between causality for query answers in databases, database repairs wrt. denial constraints, and consistency-based diagnosis. The first two are relatively new problems in databases, and the third one is an established subject of model-based diagnosis in knowledge representation.

We show how to obtain database repairs from causes, and the other way around. The vast body of research on database repairs can be applied to the newer problem of determining actual causes for query answers. By formulating a causality problem as a diagnosis problem, we manage to characterize causes in terms of the system's diagnoses. More specifically, we show that inferring and computing actual causes and responsibility in a database setting become, in different forms, consistency-based diagnosis reasoning problems and tasks.

Informally, a causal explanation for a conjunctive query answer can be viewed as a diagnosis, where in essence the first-order logical reconstruction of the relational database provides the system description [22], and the observation is the query answer. Furthermore, we unveil a strong connection between computing causes and their responsibilities for conjunctive queries, on the one hand, and computing *repairs* in databases [2] wrt. denial constraints, on the other hand. These computational problems can be reduced to each other. More precisely, we report on the following results:

1. For a boolean conjunctive query and its associated denial constraint (which is violated iff the query is true), we establish a precise connection between actual causes for the query (being true) and the subset-repairs [1] of the instance wrt. the constraint. Namely, we obtain causes from repairs.
2. In particular, we establish the connection between an actual cause's responsibility and cardinality repairs [18] wrt. the associated constraint.
3. We characterize and obtain subset- and cardinality- repairs for a database under a denial constraint in terms of the causes for the associated query being true.
4. We consider *a set* of denials constraints and a database that may be inconsistent wrt. them. We obtain the database repairs by means of an algorithm that takes as input the actual causes for constraint violations and their contingency sets.
5. We establish a precise connection between consistency-based diagnosis for a boolean conjunctive query being unexpectedly true according to a system description, and causes for the query being true. In particular, we compute actual causes, contingency sets, and responsibilities from minimal diagnosis.
6. As report on ongoing work, we discuss several extensions and open issues that are under investigation.

Acknowledgements: Leo Bertossi is grateful to Benny Kimelfeld for stimulating conversations at LogicBlox, and pointing out to [16, 17], where an interesting connection between updates through views and causality is established. He also appreciates the hospitality of LogicBlox during part of his sabbatical.

References

[1] Arenas, M., Bertossi, L. and Chomicki, J. Consistent Query Answers in Inconsistent Databases. *Proc. ACM PODS*, 1999.

[2] Bertossi, L. *Database Repairing and Consistent Query Answering.* Morgan & Claypool, Synthesis Lectures on Data Management, 2011.

[3] Borgida, A., Calvanese, D. and Rodriguez-Muro, M. Explanation in DL-Lite. *Proc. DL Workshop*, CEUR-WS 353, 2008.

[4] Buneman, P., Khanna, S. and Tan, W. C. Why and Where: A Characterization of Data Provenance. *Proc. ICDT*, 2001.

[5] Buneman, P. and Tan, W. C. Provenance in Databases. *Proc. ACM SIGMOD*, 2007.

[6] Chapman, A., and Jagadish, H. V. Why Not? *Proc. ACM SIGMOD*, 2009.

[7] Cheney, J., Chiticariu, L. and Tan, W. C. Provenance in Databases: Why, How, And Where. *Foundations and Trends in Databases*, 2009, 1(4): 379-474.

[8] Chockler, H. and Halpern, J. Y. Responsibility and Blame: A Structural-Model Approach. *J. Artif. Intell. Res.*, 2004, 22:93-115.

[9] Cui, Y., Widom, J. and Wiener, J. L. Tracing The Lineage of View Data in a Warehousing Environment. *ACM Trans. Database Syst.*, 2000, 25(2):179-227.

[10] Gertz, M. Diagnosis and Repair of Constraint Violations in Database Systems. PhD Thesis, Universität Hannover, 1996.

[11] Halpern, Y. J., and Pearl, J. Causes and Explanations: A Structural-Model Approach: Part 1 *Proc. UAI*, 2001, pp. 194-202.

[12] Halpern, Y. J., and Pearl, J. Causes and Explanations: A Structural-Model Approach: Part 1. *British J. Philosophy of Science*, 2005, 56:843-887.

[13] Huang, J., Chen, T., Doan, A. and Naughton, J. F. On The Provenance of Non-Answers to Queries over Extracted Data. *PVLDB*, 2008, 1(1):736–747.

[14] Karvounarakis, G. and Green, T. J. Semiring-Annotated Data: Queries and Provenance? *SIGMOD Record*, 2012, 41(3):5-14.

[15] Karvounarakis, G. Ives, Z. G. and Tannen, V. Querying Data Provenance. *Proc. ACM SIGMOD*, 2010, pp. 951–962.

[16] Kimelfeld, B. A Dichotomy in the Complexity of Deletion Propagation with Functional Dependencies. *Proc. ACM PODS*, 2012.

[17] Kimelfeld, B., Vondrak, J. and Williams, R. Maximizing Conjunctive Views in Deletion Propagation. *ACM Trans. Database Syst.*, 2012, 37(4):24.

[18] Lopatenko, A. and Bertossi, L. Complexity of Consistent Query Answering in Databases under Cardinality-Based and Incremental Repair Semantics. *Proc. ICDT*, 2007, Springer LNCS 4353.

[19] Meliou, A., Gatterbauer, W. Moore, K. F. and Suciu, D. The Complexity of Causality and Responsibility for Query Answers and Non-Answers. *Proc. VLDB*, 2010, pp. 34-41.

[20] Meliou, A., Gatterbauer. W., Halpern, J. Y., Koch, C., Moore K. F. and Suciu, D. Causality in Databases. *IEEE Data Eng. Bull*, 2010, 33(3):59-67.

[21] Reiter, R. A Theory of Diagnosis from First Principles. *Artificial Intelligence*, 1987, 32(1):57-95.

[22] Reiter, R. Towards a Logical Reconstruction of Relational Database Theory. In *On Conceptual Modelling*, M.L. Brodie, J. Mylopoulos and J.W. Schmidt (eds.), Springer, 1984, pp. 191-233.

[23] Tannen, V. Provenance Propagation in Complex Queries. In *Buneman Festschrift*, 2013, Springer LNCS 8000, pp. 483-493.