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

Leopoldo Bertossi and Babak Salimi

Carleton University
Ottawa, Canada
Database Causality and Goals

Causality appears at the foundations of many scientific disciplines.

Want to represent and compute causality in order to deal with the uncertainty in data and knowledge.

In data management, we need to understand and compute why certain (query) results are obtained or not.

Why certain natural semantic conditions are not satisfied.

A DB system could provide explanations.

To understand/explore the data or reconsider the query.
Our current research is motivated by trying to understand causality in data management from different perspectives.

We have established interesting and fruitful connections among four forms of reasoning:

- inferring causes from databases
- database repairs and consistent query answering (CQA)
- consistency-based and abductive diagnosis
- view updates

They all reflect some sort of uncertainty about the information at hand.

They all have related non-monotonic reasoning tasks.

(Details in Proc. NMR 2014 and Proc. BUDA (PODS WS) 2014)
A notion of causality-based explanation for a query result was introduced by (Meliou et al., VLDB 2010)

- Assume $D = D^n \cup D^x$, where $D^n$ and $D^x$ denote the sets of endogenous (candidate causes) and exogenous tuples

- A tuple $t \in D^n$ is called a counterfactual cause for a boolean conjunctive $Q$, if $D \models Q$ and $D \setminus \{t\} \not\models Q$

- A tuple $t \in D^n$ is an actual cause for $Q$ if there exists $\Gamma \subseteq D^n$, called a contingency set, such that $t$ is a counterfactual cause for $Q$ in $D \setminus \Gamma$

Based on (Halpern and Pearl, 2001, 2005)

$CS(D^n, D^x, Q)$: set of actual causes for $Q$ in $D = D^n \cup D^x$
Responsibility reflects relative degree of causality of a tuple for a query result (Meliou et al., VLDB 2010)

- The responsibility of an actual cause $t$ for $Q$:

$$\rho_D(t) := \frac{1}{|\Gamma| + 1}$$

$|\Gamma|$ = size of smallest contingency set for $t$

Tuples with higher responsibility tend to provide more interesting explanations for query results

Based on (Chockler and Halpern, 2004)
**Example:** Database $D$ with relations $R$ and $S$ below

$Q: \exists x \exists y (S(x) \land R(x, y) \land S(y))$ \hspace{1cm} $D \models Q$

Causes for $Q$ being true in $D$ (assume all tuples endogenous):

<table>
<thead>
<tr>
<th>$R$</th>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_4$</td>
<td>$a_3$</td>
<td></td>
</tr>
<tr>
<td>$a_2$</td>
<td>$a_1$</td>
<td></td>
</tr>
<tr>
<td>$a_3$</td>
<td>$a_3$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$S$</th>
<th>$X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_4$</td>
<td></td>
</tr>
<tr>
<td>$a_2$</td>
<td></td>
</tr>
<tr>
<td>$a_3$</td>
<td></td>
</tr>
</tbody>
</table>

$S(a_3)$ is counterfactual cause for $Q$: if $S(a_3)$ is removed from $D$, $Q$ is no longer an answer; its responsibility is 1

$R(a_4, a_3)$ is an actual cause for $Q$ with contingency set \{ $R(a_3, a_3)$ \}

If $R(a_3, a_3)$ is removed from $D$, $Q$ is still true, but further removing $R(a_4, a_3)$ makes $Q$ false

The responsibility of $R(a_4, a_3)$ is $\frac{1}{2}$: its smallest contingency sets have size 1

$R(a_3, a_3)$ and $S(a_4)$ are actual causes, with responsibility $\frac{1}{2}$
Database Repairs

Example: Denial constraints (DC) and inconsistent database

\[ \neg P(x), Q(x, y) \]
\[ \neg P(x), R(x, y) \]

Minimal subset-repairs (S-repairs):

\[ D_1 = \{ P(e), Q(a, b), R(a, b) \} \quad D_2 = \{ P(e), P(a) \} \]

Minimal cardinality-repairs (C-repairs): \[ D_1 \]
Causes from Repairs

Assume BCQ: \( Q : \exists \bar{x} (P_1(\bar{x}_1) \land \cdots \land P_m(\bar{x}_m)) \)

\( Q \) is unexpectedly true in \( D \) (or we expected \( D \models \neg Q \))

What are the causes for \( Q \) to be true?

We can obtain actual causes and contingency sets from database repairs

\( \neg Q \) logically equivalent to DC:

\( \kappa(Q) : \forall \bar{x} \neg (P_1(\bar{x}_1) \land \cdots \land P_m(\bar{x}_m)) \)

Also as the Datalog constraint: \((\bot) \leftarrow P_1(\bar{x}_1), \ldots, P_m(\bar{x}_m)\)

If \( Q \) not expected to hold, we may consider \( D \) to be inconsistent wrt \( \kappa(Q) \)
Repairs of $D$ wrt $\kappa(Q)$ may be considered ...

First consider class $S\text{rep}(D, \kappa(Q))$ of $S$-repairs

With DCs, $S$-repairs are $S$-maximally consistent subsets of $D$ (no proper subset that is a repair)

Collect set differences between $D$ and those $S$-repairs that do not contain tuple $t \in D^n$, and are obtained by removing endogenous tuples:

$$\mathcal{DF}(D, D^n, \kappa(Q), t) = \{ D \setminus D' \mid D' \in S\text{rep}(D, \kappa(Q)) \},$$

**Proposition:** $D = D^n \cup D^x$, $t \in D^n$ \quad $t \in (D \setminus D') \subseteq D^n$

(a) $t$ is an actual cause for $Q$ iff $\mathcal{DF}(D, D^n, \kappa(Q), t) \neq \emptyset$

(b) If $\mathcal{DF}(D, D^n, \kappa(Q), t) = \emptyset$, then $\rho_D(t) = 0$

(c) If $\rho_D(t) \neq 0$, $\rho_D(t) = \frac{1}{|s|}$, where $s \in \mathcal{DF}(D, D^n, \kappa(Q), t)$ and there is no $s' \in \mathcal{DF}(D, D^n, \kappa(Q), t)$ with $|s'| < |s|$
Repairs from Causes

Database instance $D$ and DC $\kappa: \leftarrow A_1(\bar{x}_1), \ldots, A_n(\bar{x}_n)$

Boolean conjunctive violation view can be associated to $\kappa$:

$$V^\kappa: \exists \bar{x}(A_1(\bar{x}_1) \land \cdots \land A_n(\bar{x}_n))$$

If $D$ inconsistent wrt $\kappa \Rightarrow$ query $V^\kappa$ becomes true

Use actual causes for $V^\kappa$ to obtain $S$-repairs of $D$ wrt $\kappa$ ...

Collect all $S$-minimal contingency sets associated with actual cause $t$ for $V^\kappa$:

$CT(D, D^n, V^\kappa, t)$: collects all subsets $s \subseteq D^n$ with

(a) $D \setminus s \models V^\kappa$

(b) $D \setminus (s \cup \{t\}) \not\models V^\kappa$

(c) $\forall s'' \subsetneq s: D \setminus (s'' \cup \{t\}) \models V^\kappa$
Proposition: (here all tuples endogenous)

(a) \( D \) is consistent wrt. \( \kappa \) iff \( \mathcal{CS}(D, \emptyset, V^\kappa) = \emptyset \)

(b) \( D' \subseteq D \) is an S-repair for \( D \) iff, for every \( t \in D \setminus D' \),
\( t \in \mathcal{CS}(D, \emptyset, V^\kappa) \) and \( D \setminus (D' \cup \{t\}) \in \mathcal{CT}(D, D, V^\kappa, t) \)

And CQA?

Instance \( D \), and a DC \( \kappa \)

A relationship between consistent query answering (CQA) wrt the S-repair semantics and actual cases for the violation view \( V^\kappa \)

Proposition: Ground atomic query \( A \) is \textit{Srep}-consistently true
iff \( A \in D \setminus \mathcal{CS}(D, \emptyset, V^\kappa) \)
It is possible to obtain repairs from causes for (violations of) sets of DCs

Roughly ...

To obtain S-repairs:

- Form set of all actual causes for violation views:
  \[ S = \{ t \mid \exists \psi \in \Sigma: \ t \in CS(D,V^\psi) \} \]

- Form collection of non-empty sets of actual causes for each violation view:
  \[ C = \{ CS(D,V^\psi) \mid \exists \psi \in \Sigma, CS(D,V^\psi) \neq \emptyset \} \]

- \( C \) is a collection of subsets of set \( S \)

- Compute S-minimal \textit{hitting sets} of collection \( C \)

- Build repairs using the hitting sets ...
We can also obtain C-repairs from actual causes ...

C-repairs are related to most responsible actual causes ...

$\mathcal{MRC}(D, V^\kappa)$ collects most responsible actual causes $t \in D$ for $V^\kappa$, i.e.

(a) $t \in \mathcal{CS}(D, \emptyset, V^\kappa)$

(b) There is no $t' \in \mathcal{CS}(D, \emptyset, V^\kappa)$ with $\rho_D(t') > \rho_D(t)$

**Proposition:** For $D$ and $DC^\kappa$, $D'$ is a C-repair for $D$ wrt $\kappa$ iff for each $t \in D \setminus D'$: $t \in \mathcal{MRC}(D, V^\kappa)$ and $D \setminus (D' \cup \{t\}) \in \mathcal{T}(D, D, V^\kappa, t)$
Consistency-Based Diagnosis and DB Causality

\[ D = D^n \cup D^x \] instance for schema \( S \)

**Example:** \( D = \{ S(a_3), S(a_4), R(a_4, a_3) \}, \ D^n = \{ S(a_4), S(a_3) \} \)

\( Q : \exists x \exists y (S(x) \land R(x, y) \land S(y)) \)

**System description:** \( \mathcal{M} = (SD, \{ S(a_4), S(a_3) \}, Q) \)

\( SD \) contains:
1. Reiter’s logical reconstruction of \( D \):

   (a) **Predicate completion** axioms:
   
   \[ \forall xy (R(x, y) \iff x = a_4 \land y = a_3), \]
   
   \[ \forall x (S(x) \iff x = a_3 \lor x = a_4) \]

   (b) **Unique names assumption:** \( a_4 \neq a_3 \)
2. “Normal” tuples satisfy the DC

\[(c) \ \kappa(Q)^{ext} : \forall xy \neg(S(x) \land \neg ab S(x) \land R(x, y) \land \neg ab R(x, y) \land S(y) \land \neg ab S(y))\]

3. Relationship between query and original DC and inclusion dependencies

\[(d) \ \neg \kappa(Q) \iff Q \quad (\kappa(Q) \text{ and } Q \text{ as before})\]

\[(e) \ \forall xy(ab R(x, y) \to R(x, y)), \ \forall x(ab S(x) \to S(x))\]

If we put together \textit{SD}, the observation \(Q\), and the assumption about the normality of tuples, i.e.

\[
\forall xy(ab R(x, y) \to \text{false}), \ \forall x(ab S(x) \to \text{false})
\]

We obtain an inconsistent FO theory!

From it we can obtain diagnoses ...
A diagnosis for $\mathcal{M}$ is $\Delta \subseteq D^n$:

$$SD \cup \{ab_P(\bar{c}) \mid P(\bar{c}) \in \Delta\} \cup \{\neg ab_P(\bar{c}) \mid P(\bar{c}) \in D \setminus \Delta\} \cup \{Q\}$$

is consistent

$\mathcal{D}(\mathcal{M}, t)$: class of all S-minimal diagnoses for $\mathcal{M}$ that contain tuple $t \in D^n$

$\mathcal{MCD}(\mathcal{M}, t)$: class of diagnoses of $\mathcal{M}$ that contain tuple $t \in D^n$ and have the minimum cardinality (among those diagnoses that contain $t$)

Clearly: $\mathcal{MCD}(\mathcal{M}, t) \subseteq \mathcal{D}(\mathcal{M}, t)$

Proposition:

(a) Tuple $t \in D^n$ is an actual cause for $Q$ iff $\mathcal{D}(\mathcal{M}, t) \neq \emptyset$

(b) For tuple $t \in D^n$, $\rho_D(t) = 0$ iff $\mathcal{MCD}(\mathcal{M}, t) = \emptyset$

Otherwise, $\rho_D(t) = \frac{1}{|s|}$, where $s \in \mathcal{MCD}(\mathcal{M}, t)$
Conclusions

• It is possible to take advantage of results and techniques for
  • database repairs and CQA
  • consistency-based diagnosis
    (Reiter; 87) uses hitting sets as an algorithmic basis
to obtain algorithms and complexity results for causality and responsibility problems for conjunctive queries

Coming in new paper ...

• The partition of a database into endogenous and exogenous tuples has been used in causality

This kind of partition is also of interest in the context of repairs
• Considering that we should have more control on endogenous tuples than on exogenous ones, which may come from external sources, it makes sense to:

For example, consider *endogenous repairs* that are obtained by updates (of any kind) on endogenous tuples

For example, in the case of violation of denial constraints, endogenous repairs would be obtained -if possible- by deleting endogenous tuples only

• If there are no repairs based on endogenous tuples only, a preference condition could be imposed on repairs

Privilege those that change exogenous the least (or the other way around)
• We could go even further and apply notions of *preferred repairs* (Chomicki et al., 2012)

Assume $\text{Prep}$ denotes a given class of *preferred repairs*

We can use relationships as above, replacing $\text{Srep}$ by $\text{Prep}$, and define other notions of *preferred causes*

• As a further extension, we could assume that combinations of (only) exogenous tuples never violate the ICs (could be checked at upload time)

That a part of the database could be considered to be consistent, while the other is subject to possible repairs

A situation like this has been considered, for other purposes and in a different form, in (Greco et al., VLDB 2014)