Tutorial: Provenance and Causality

Wolfgang Gatterbauer, Alexandra Meliou, Dan Suciu
Overview

• Let $Q(D) = D'$ be a database transformation
• Let $t' \in D'$ be an output tuple
• Which tuples $t \in D$ caused $t' \in D'$?

• This talk: review causality, define causality in database transformation, give applications
Credits

- *Causality*, Judea Pearl, 2000
- *Complexity results for explanations in the structural-model approach*, Eiter and Lukasiewicz, 2004
- *Scalable Techniques for Mining Causal Structures*, Silverstein, Brin, Motwani, Ullman, 2000
- Responsibility and blame: A structural-model approach, Chockler, Halpern, 2004
- Y. Crama, P. L. Hammer, Boolean Functions: Theory, Algorithms, and


- *Causality in Databases*, Meliou et al., 2010
- *Tracing Data Errors with View-Conditioned Causality*, Meliou et al., 2011
- *The Complexity of Causality and Responsibility for Query Answers and non-Answers*, Meliou et al., 2011b
- WHY SO? or WHY NO? Functional Causality for Explaining Query Answers, Meliou, et al., 2010
- *Causality and the Semantics of Provenance*, James Cheney, DCM 2010
Outline

• Brief History
• Causality for a Boolean function
• Causality for a query
• Applications
• Summary
History of Causality

• What is the mathematical equation of cause?

• Surprisingly difficult to give

• The following brief “history” is based mostly on [Pearl’2000] – an advertisement for this great book
Antiquity

- Causality used to pass responsibility, attributing intent, and blame: only gods, humans, animals are agents of cause

- Aristotle viewed causes in terms of a purpose; no definition
The Dawn of Science & Engineering

• Find objective causes rather than passing responsibility

• Questions of interest:
  – **Why** doesn’t the wheel turn?
  – **What if** I make the beam half as thick, will it carry the load?
  – **How** do I shape the beam so it will carry the load?

• Same questions today in DB!

**But what IS causality?**
Science Seeks to Explain Causes, but *Lacks* A Language For Causality

Second law of motion says this:

$$F = m \cdot a$$

• What we know, but the law doesn’t say it: The force *causes* acceleration: $$a = F/m$$

• We also know, but the law doesn’t say it: Force + acceleration *determine* mass: $$m = F/a$$ They do *not* *cause* the mass.
David Hume

• Causality is a matter of perception:
  – “we remember seeing the flame, and feeling a sensation called heat; without farther ceremony, we call the one cause and the other effect”.
• Opens door to finding causes from empirical observations
• But correlation is not causation
Karl Pearson

• Co-founder of modern statistics
• Forget causation! Correlation is all you should ask for
• Statistical machine learning (ML) relies on the principle of seeking only correlations
• Mantra: don’t attempt to find causation!
• Very few dissenters dare to look for causations in data, e.g. [Silverstein’2000]
Judea Pearl

- Forget empirical observations
- Start from a **Causal Network**, consisting of known, physical causation relationships
- Substitute randomness with **Exogenous Variables**

- Result: a mathematical definition of causality
- Caveat: must build the causal network first

In database transformations:
causal network = provenance
Outline

• Brief History
• Causality for a Boolean function
• Causality for a query
• Applications
• Summary
Causal Model

**Definition** A *causal model* (causal network) consists of:

- **Exogenous** Boolean variables -- external; fixed
- **Endogenous** Boolean variables -- internal; modifiable
- Boolean functions defining some endogenous variables

This talk: Causal Model is a Boolean function $Y = F(X_1, \ldots, X_n)$
Cause of a Boolean Function $Y = F(X_1,...,X_n)$

- Fix $\theta = \text{an assignment (world)}; \quad y = \theta(F)$
- Fix $X_i$ (endogenous variable); \quad $x_i = \theta(X_i)$
- $[\theta-X_i](F) = \text{assign all variables in } F \text{ except } X_i$

**Definition** The event $X_i = x_i$ is a **counterfactual cause** for $Y = y$ in $\theta$, if $[\theta-X_i](F)$ depends on $X_i$

Equivalently: changing the value of $X_i$ in $\theta$ causes $Y$ to change

**Definition** The event $X_i = x_i$ is an **actual cause** for $Y = y$ if $\exists \theta'$ s.t. it is counterfactual for $Y = y$ in $\theta'$

$\theta$, $\theta'$ must agree on $X_i$, on exogenous variables, on output $Y$
Three Simple Examples

Assume all variables $X_1$, $X_2$, $X_3$ are endogenous

1. $Y = X_1 \land X_2$
   - $\theta(X_1) = \theta(X_2) = 1 \implies Y = 1$
   - $[\theta-X_1](F) = X_1 \land 1 = X_1$
   - $X_1 = 1$ is a counterfactual cause for $Y = 1$

2. $Y = X_1 \lor X_2$
   - $\theta(X_1) = \theta(X_2) = 1 \implies Y = 1$
   - $[\theta-X_1](F) = X_1 \lor 1 = 1$
   - $X_1 = 1$ is no counterfactual cause for $Y = 1$

3. $Y = [\text{not}(X_1) \land X_2] \lor X_3$
   - $\theta'(X_1) = 1, \theta'(X_2) = 0 \implies [\theta'-X_1](F) = X_1$
   - $X_1 = 1$ is an actual cause for $Y = 1$
   - $\theta(X_1) = \theta(X_2) = \theta(X_2) = 1 \implies Y = 1$
   - $[\theta'-X_1](F) = \text{not}(X_1)$

   $X_1 = 1$ is not an actual cause for $Y = 1$
Complexity of Causality

[Eiter&Lukasiewicz 2004]

<table>
<thead>
<tr>
<th>Counterfactual cause</th>
<th>Actual cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTIME</td>
<td>NP-complete</td>
</tr>
</tbody>
</table>

**Proof:** Reduction from SAT.
Given $F$, $F$ is satisfiable iff $X$ is an actual cause for $X \land F$
Related Concepts 1/3

**Definition** Fix $F$, $\theta$. $X_i$ is a **critical** for $F$ in $\theta$ if $[\theta - X_i](F)$ depends on $X_i$

“$X_i$ is counterfactual cause for $F$ in $\theta$”
“$X_i$ is a critical voter (swing voter)”

**Definition** Fix $F$. $X_i$ is a **critical** for $F$ if $\exists \theta$ s.t $X_i$ is critical for $F$ in $\theta$

“$F$ depends on $X_i$” “$X_i$ is in the support of $F$”

Applications to data privacy:
if $X_i$ is not critical, then $F$ reveals nothing about $X_i$. 
Related Concepts 2/3

**Definition** Fix F.

The influence of $X_i$: $\text{Inf}(X_i) = \text{Prob}[X_i \text{ is critical for } F, \theta]$

**Examples:**
- Majority function: $\text{Inf}(X_i) = \left(\frac{n-1}{2}\right) \frac{1}{2^{n-1}} \approx \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{n}}$
- Parity function: $\text{Inf}(X_i) = 1$

**Application:** The influence (or power) of voter $X_i$

Modified Chow Index = $\text{Inf}(X_i) \times 2^n$

Banzhaf Index = $\frac{\text{Inf}(X_i)}{[\text{Inf}(X_1) + ... + \text{Inf}(X_n)]}$

[O’Donnell 2008]

Probability over random choices of $\theta$

[Crama’2010 p.87 and p78]
$X_i$ is counterfactual for $F$ in $\theta$ $\iff$ $X_i$ is critical for $F$ in $\theta$ $\iff$ $\text{Inf}(X_i) > 0$ $\iff$ $X_i$ is critical for $F$
Outline

• Brief History
• Causality for a Boolean function
• Causality for a query
• Applications
• Summary
Endogenous/exogenous Tuples

Input database $D$, Query $Q$, output $D' = Q(D)$

Two kinds of tuples can go into $D$:
- **Exogenous tuples**: $D^x$
  - External, from sources that are certain; not causes
- **Endogenous tuples**: $D^n$
  - Tuples that affect outcome; potential causes

Database $D \subseteq D^x \cup D^n$
Causality of a Query Answer

Fix database $D$, query $Q$, output $D' = Q(D)$

- Input $t \in D^n$ is **counterfactual cause in $D$ for $t'$**
  if $t'$ occurs in exactly one of $Q(D)$ or $Q(D \otimes t)$
  - **Why-so cause**: when $t' \in Q(D)$ and $t' \notin Q(D \otimes t)$
  - **Why-no cause**: when $t' \notin Q(D)$ and $t' \in Q(D \otimes t)$

- Input $t \in D^n$ is an **actual cause in $D$ for output $t'$**
  if $\exists \Gamma \subseteq D^n$ s.t. $t$ is a counterfactual cause in $D \otimes \Gamma$

**Contingency set**

**symmetric difference**
Responsibility of a Query Answer

**Definition:** Responsibility of \( t \) for \( t' \)

\[
\rho_t = \frac{1}{1 + \min_{\Gamma} |\Gamma|}
\]

Here \( \Gamma \subseteq D^n \) ranges over contingency sets

- If \( \rho_t = 1 \) then \( t \) is a **counterfactual cause**
- If \( 0 < \rho_t < 1 \) then \( t \) is an **actual cause**
- If \( \rho_t = 0 \) then \( t \) is **not a cause**

Responsibility introduced for *causal networks* [Chockler’2004]
Complexity

### Causality of CQ queries;

<table>
<thead>
<tr>
<th>Why-so?</th>
<th>Why-no?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTIME</td>
<td>PTIME</td>
</tr>
</tbody>
</table>

### Responsibility of CQ queries without self-joins;

<table>
<thead>
<tr>
<th>Why-so?</th>
<th>Why-no?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTIME/NP-hard dichotomy</td>
<td>PTIME</td>
</tr>
</tbody>
</table>

Some queries in PTIME are NL-hard (hence not in FO)

Related: hardness of responsibility in causal networks [Chockler’2004]
Responsibility: Dichotomy

**Theorem** Data complexity of the responsibility:
- If $Q$ is *weakly linear*, then $Q$ is in PTIME
- If $Q$ is not *weakly linear*, then it is NP-hard

See [Meliou’2011b] for a definition of “weakly linear”.

Will give examples next.
Responsibility: Easy and Hard Queries

Example Responsibility for the following query is in PTIME

\[ Q : - R(x, y), S(y, z), T(z, u), M(u, v), \ldots \]

Example: Responsibility for these queries is NP-hard:

\[ h_1^* : - A^n(x), B^n(y), C^n(z), W(x, y, z) \]
\[ h_2^* : - R^n(x, y), S^n(y, z), T^n(z, x) \]
\[ h_3^* : - A^n(x), B^n(y), C^n(z), R(x, y), S(y, z), T(z, x) \]
Related Concept

The *Deletion Propagation Problem* [Buneman’2002]

- Fix $D$, $Q$, and $t' \subseteq Q(D)$
- Problem: find $\Gamma \subseteq D$ such that
  - $t' \not\in Q(D - \Gamma)$
  - The side effects $|Q(D) \otimes Q(D - \Gamma)|$ are minimal

[Kimelfeld’2011] prove a dichotomy into PTIME and NP-hard for the Deletion Propagation Problem

Intuitively, $\Gamma$ acts like a contingency set, but precise connection to causality has not been studied
Outline

• Brief History
• Causality for a Boolean function
• Causality for a query
• Applications
• Summary
One or more outputs are wrong. Which inputs need to be corrected? “Post factum” data cleaning.
Example

Input Data

- Accelerometer
- GPS
- Cell Tower
- Audio
- Light

Transformations

- Periodicity $p$
- Has Signal? $h$
- Speed $s$
- Rate of Change $r$
- Avg. Strength $a$
- Zero crossing rate $z$
- Spectral roll-off $c$
- Avg. Intensity $i$

Output Data

- Is Walking? $M(p > P_w, R_s < r < R_w, ¬h ∨ (s < S_w))$
- Is Driving? $M(p < P_d, r > R_d, h, s > S_d)$
- Alone? $(A_2 ≥ a > A_1) ∨ ((a > A_2) ∧ (z > Z)) ∨ ((a > A_3) ∧ (z < Z) ∧ (c > C))$
- Is Indoor? $M(¬h, i < I_i)$
- Is Meeting? $M(¬h, i < I_m, a > A_m, z > Z_m)$

Table:

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
<th>Value 6</th>
<th>Value 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerometer</td>
<td>0.016</td>
<td>True</td>
<td>0.067</td>
<td>0</td>
<td>0.4</td>
<td>0.004</td>
<td>0.86</td>
</tr>
<tr>
<td>GPS</td>
<td>0.0009</td>
<td>False</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0.0039</td>
<td>0.81</td>
</tr>
<tr>
<td>Cell Tower</td>
<td>0.005</td>
<td>True</td>
<td>0.19</td>
<td>0</td>
<td>0.03</td>
<td>0.003</td>
<td>0.75</td>
</tr>
<tr>
<td>Audio</td>
<td>0.0008</td>
<td>True</td>
<td>0.003</td>
<td>0</td>
<td>0.1</td>
<td>0.003</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Sensors may be faulty or inhibited. It is not straightforward to spot such errors in the provenance.

What caused these errors?

[Source: Meliou'2011]
Average precision is a metric of quality of a ranking.

If all erroneous variables are ranked first, then average precision is 1.

800 different instances
5 sensory inputs
8 extracted features (variables)
3 users
~10% observed errors

Static analysis of lineage

Simpler causality schemes
We select the highest responsibility variable, remove it from the evaluation of all classifiers, and record the portion of errors that get corrected per classifier.

Driving has reliable features (low responsibility), means they are almost never causes of error.

Walking has no reliable features.

We achieve almost 90% correction ratio for "driving"!

But we can only fix few "walking" errors (?)
Why-So / Why-No Queries

Database schema

Director($did$, $firstName$, $lastName$)
Movie($mid$, $name$, $year$, $rank$)
Movie_Directors($did$, $mid$)
Genre($mid$, $genre$)

Query

SELECT DISTINCT g.genre
FROM Director d, Movie_Directors md,
    Movie m, Genre g
WHERE d.lastName LIKE ‘Burton’
AND g.mid=m.mid
AND m.mid=md.mid
AND md.did=d.did
ORDER BY g.genre

Query answers

<table>
<thead>
<tr>
<th>genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>…</td>
</tr>
<tr>
<td>Drama</td>
</tr>
<tr>
<td>Family</td>
</tr>
<tr>
<td>Fantasy</td>
</tr>
<tr>
<td>History</td>
</tr>
<tr>
<td>Horror</td>
</tr>
<tr>
<td>Music</td>
</tr>
<tr>
<td>Musical</td>
</tr>
<tr>
<td>Mystery</td>
</tr>
<tr>
<td>Romance</td>
</tr>
<tr>
<td>Sci-Fi</td>
</tr>
<tr>
<td>…</td>
</tr>
</tbody>
</table>
query answer: 

<table>
<thead>
<tr>
<th>Score</th>
<th>Movie(526338, &quot;Sweeney Todd&quot;, 2007)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33</td>
<td>Director(23456, David, Burton)</td>
</tr>
<tr>
<td>0.33</td>
<td>Director(23468, Humphrey, Burton)</td>
</tr>
<tr>
<td>0.33</td>
<td>Director(23488, Tim, Burton)</td>
</tr>
<tr>
<td>0.25</td>
<td>Movie(359516, &quot;Let's Fall in Love&quot;, 1933)</td>
</tr>
<tr>
<td>0.25</td>
<td>Movie(565577, &quot;The Melody Lingers On&quot;, 1935)</td>
</tr>
<tr>
<td>0.20</td>
<td>Movie(6539, &quot;Candide&quot;, 1989)</td>
</tr>
<tr>
<td>0.20</td>
<td>Movie(173629, &quot;Flight&quot;, 1999)</td>
</tr>
<tr>
<td>0.20</td>
<td>Movie(389987, &quot;Manon Lescaut&quot;, 1997)</td>
</tr>
</tbody>
</table>
Outline

• Brief History
• Causality for a Boolean function
• Causality for a query
• Applications
• Summary
Summary

• Causality in Data Transformations:
  – Which input tuple caused this output tuple?

• Key concepts:
  – Endogenous/exogenous tuples
  – Counterfactual cause
  – Contingency set
  – Actual cause

• Very simple causal network: Boolean function
  – Avoids many complications

• Applications:
  – Error corrections
  – Why-so / why-no explanations