Query-Answer Causality in Databases: Abductive Diagnosis and View-Updates

Babak Salimi & Leopoldo Bertossi
Carleton University
School of Computer Science
Ottawa, Canada

Database Causality and Goals

- Causality appears at the foundations of many scientific disciplines

Database Causality and Goals

- Causality appears at the foundations of many scientific disciplines
- Want to represent and compute causality, to deal with the *uncertainty* in data and knowledge

Database Causality and Goals

- Causality appears at the foundations of many scientific disciplines
- Want to represent and compute causality, 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

Database Causality and Goals

- Causality appears at the foundations of many scientific disciplines
- Want to represent and compute causality, 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

Database Causality and Goals

- Causality appears at the foundations of many scientific disciplines
- Want to represent and compute causality, 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 should provide *explanations*

Database Causality and Goals

- Causality appears at the foundations of many scientific disciplines
- Want to represent and compute causality, 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 should provide *explanations*
- To understand/explore the data or reconsider the query

Database Causality and Goals

- Causality appears at the foundations of many scientific disciplines
- Want to represent and compute causality, 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 should provide *explanations*
- To understand/explore the data or reconsider the query
- Enhancing the value of extracted data

Database Causality and Goals

- Causality appears at the foundations of many scientific disciplines
- Want to represent and compute causality, 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 should provide *explanations*
- To understand/explore the data or reconsider the query
- Enhancing the value of extracted data
- Crucial for understanding massive volumes of data!

Our Research Motivation

- Our current research is motivated by trying to understand causality in data management from different perspectives

Our Research Motivation

- Our current research is motivated by trying to understand causality in data management from different perspectives
- We have already established interesting and fruitful connections among three forms of reasoning: (Salimi & Bertossi, ICDT 2015)

Our Research Motivation

- Our current research is motivated by trying to understand causality in data management from different perspectives
- We have already established interesting and fruitful connections among three forms of reasoning: (Salimi & Bertossi, ICDT 2015)
  - inferring causes from databases

Our Research Motivation

- Our current research is motivated by trying to understand causality in data management from different perspectives
- We have already established interesting and fruitful connections among three forms of reasoning: (Salimi & Bertossi, ICDT 2015)
  - inferring causes from databases
  - database repairs and consistent query answering (CQA)

Our Research Motivation

- Our current research is motivated by trying to understand causality in data management from different perspectives
- We have already established interesting and fruitful connections among three forms of reasoning: (Salimi & Bertossi, ICDT 2015)
  - inferring causes from databases
  - database repairs and consistent query answering (CQA)
  - consistency-based diagnoses

Our Research Motivation

- Our current research is motivated by trying to understand causality in data management from different perspectives
- We have already established interesting and fruitful connections among three forms of reasoning: (Salimi & Bertossi, ICDT 2015)
  - inferring causes from databases
  - database repairs and consistent query answering (CQA)
  - consistency-based diagnoses
- In this work we add:

Our Research Motivation

- Our current research is motivated by trying to understand causality in data management from different perspectives
- We have already established interesting and fruitful connections among three forms of reasoning: (Salimi & Bertossi, ICDT 2015)
  - inferring causes from databases
  - database repairs and consistent query answering (CQA)
  - consistency-based diagnoses
- In this work we add:
  - view updates (database updates through views)

Our Research Motivation

- Our current research is motivated by trying to understand causality in data management from different perspectives
- We have already established interesting and fruitful connections among three forms of reasoning: (Salimi & Bertossi, ICDT 2015)
  - inferring causes from databases
  - database repairs and consistent query answering (CQA)
  - consistency-based diagnoses
- In this work we add:
  - view updates (database updates through views)
  - abductive diagnosis

Commonalities of the Problems

- Those reasoning problems share some commonalities:

Commonalities of the Problems

- Those reasoning problems share some commonalities:
  - They reflect some sort of *uncertain information*

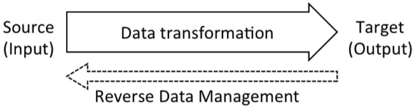
Commonalities of the Problems

- Those reasoning problems share some commonalities:
  - They reflect some sort of *uncertain information*
  - They are *non-monotonic*

Commonalities of the Problems

- Those reasoning problems share some commonalities:
  - They reflect some sort of *uncertain information*
  - They are *non-monotonic*
  - They are forms of *reverse data transformation*

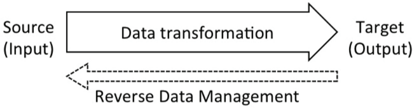
    To reason "diagnostically" about an output (observation)



Commonalities of the Problems

- Those reasoning problems share some commonalities:
  - They reflect some sort of *uncertain information*
  - They are *non-monotonic*
  - They are forms of *reverse data transformation*

    To reason "diagnostically" about an output (observation)



  - These problems have been classified under *reverse data management* (Meliou et al., VLDB 2011)

Actual Causality

A notion of causality-based explanation for a query result was introduced in (Meliou et al., VLDB 2010)

Actual Causality

A notion of causality-based explanation for a query result was introduced in (Meliou et al., VLDB 2010)

- Assume $D = D^n \cup D^x$
  D^n endogenous tuples (candidate causes)
  D^x exogenous tuples

Actual Causality

A notion of causality-based explanation for a query result was introduced in (Meliou et al., VLDB 2010)

- Assume $D = D^n \cup D^x$
  D^n endogenous tuples (candidate causes)
  D^x exogenous tuples

- $\tau \in D^n$ is counterfactual cause for answer $\bar{a}$ to monotone query $\mathcal{Q}(\bar{x})$ in D, if $D \models \mathcal{Q}(\bar{a})$ and $D \smallsetminus \{\tau\} \not\models \mathcal{Q}(\bar{a})$

Actual Causality

A notion of causality-based explanation for a query result was introduced in (Meliou et al., VLDB 2010)

- Assume $D = D^n \cup D^x$
  D^n endogenous tuples (candidate causes)
  D^x exogenous tuples
- $\tau \in D^n$ is counterfactual cause for answer $\bar{a}$ to monotone query $\mathcal{Q}(\bar{x})$ in D, if $D \models \mathcal{Q}(\bar{a})$ and $D \smallsetminus \{\tau\} \not\models \mathcal{Q}(\bar{a})$
- $\tau \in D^n$ is an actual cause for $\bar{a}$ if there is $\Gamma \subseteq D^n$, a contingency set, such that τ is a counterfactual cause for $\bar{a}$ in $D \smallsetminus \Gamma$

Actual Causality

A notion of causality-based explanation for a query result was introduced in (Meliou et al., VLDB 2010)

- Assume $D = D^n \cup D^x$
  D^n endogenous tuples (candidate causes)
  D^x exogenous tuples

- $\tau \in D^n$ is counterfactual cause for answer $\bar{a}$ to monotone query $\mathcal{Q}(\bar{x})$ in D, if $D \models \mathcal{Q}(\bar{a})$ and $D \smallsetminus \{\tau\} \not\models \mathcal{Q}(\bar{a})$

- $\tau \in D^n$ is an actual cause for $\bar{a}$ if there is $\Gamma \subseteq D^n$, a contingency set, such that τ is a counterfactual cause for $\bar{a}$ in $D \smallsetminus \Gamma$

Based on Halpern & Pearl's actual causation w/deletion as "intervention" (Halpern & Pearl, 2001, 2005)

Causal Responsibility

Causal responsibility reflects *relative degree of causality* of a tuple for a query result (Meliou et al., VLDB 2010)

- The responsibility of an actual cause τ for $\bar{a}$:

$$\rho_{D,\mathcal{Q}(\bar{a})}(\tau) = \frac{1}{|\Gamma|+1}$$

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

Causal Responsibility

Causal responsibility reflects *relative degree of causality* of a tuple for a query result (Meliou et al., VLDB 2010)

- The responsibility of an actual cause τ for $\bar{a}$:

$$\rho_{D,\mathcal{Q}(\bar{a})}(\tau) = \frac{1}{|\Gamma|+1}$$

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

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

Based on (Chockler and Halpern, 2004)

Causality and Responsibility for Query Results

Example: Database D

Author	*AuName*	*Journal*
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	*Journal*	*Topic*	*#Paper*
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

Causality and Responsibility for Query Results

Example: Database D

Author	*AuName*	*Journal*
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	*Journal*	*Topic*	*#Paper*
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

Conjunctive query $\mathcal{Q}(x, y)$:

$$\begin{aligned} Ans(AuName, Topic) \longleftarrow \; & Author(AuName, Journal), \\ & Journal(Journal, Topic, \#Paper), \end{aligned}$$

Causality and Responsibility for Query Results

Example: Database D

Author	AuName	Journal
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	Journal	Topic	#Paper
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

Conjunctive query $\mathcal{Q}(x, y)$:

$$Ans(AuName, Topic) \longleftarrow Author(AuName, Journal), \; Journal(Journal, Topic, \#Paper),$$

Answers:

$\mathcal{Q}(D)$	AuName	Topic
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

Causality and Responsibility for Query Results

Example: (cont.)

(John, XML): unexpected answer

What are the causes ?

Author	AuName	Journal
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	Journal	Topic	#Paper
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$$\mathit{Ans}_{\mathcal{Q}}(\mathit{AuName}, \mathit{Topic}) \longleftarrow \mathit{Author}(\mathit{AuName}, \mathit{Journal}), \mathit{Journal}(\mathit{Journal}, \mathit{Topic}, \#\mathit{Paper}),$$

Answers:

$\mathcal{Q}(D)$	AuName	Topic
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

Causality and Responsibility for Query Results

Example: (cont.)

Author	*AuName*	*Journal*
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	*Journal*	*Topic*	*#Paper*
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$Ans_{\mathcal{Q}}(AuName, Topic) \longleftarrow Author(AuName, Journal), Journal(Journal, Topic, \#Paper),$

Answers:

$\mathcal{Q}(D)$	*AuName*	*Topic*
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

(John, XML): unexpected answer

What are the causes ?

$\tau =$ *Author(John, TODS)* is an actual cause, with contingency sets

$\Gamma_1 = \{$ *Author(John, TKDE)* $\}$ and

$\Gamma_2 = \{$*Journal(TKDE, XML, 30)*$\}$

Causality and Responsibility for Query Results

Example: (cont.)

Author	AuName	Journal
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	Journal	Topic	#Paper
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$$Ans_{\mathcal{Q}}(AuName, Topic) \longleftarrow Author(AuName, Journal), Journal(Journal, Topic, \#Paper),$$

Answers:

$\mathcal{Q}(D)$	AuName	Topic
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

(John, XML): unexpected answer

What are the causes ?

$\tau =$ *Author(John, TODS)* is an actual cause, with contingency sets

$\Gamma_1 = \{$ *Author(John, TKDE)* $\}$ and

$\Gamma_2 = \{$*Journal(TKDE, XML, 30)*$\}$

$\rho(\tau) = \frac{1}{2}$

Causality and Responsibility for Query Results

Example: (cont.)

Author	*AuName*	*Journal*
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	*Journal*	*Topic*	*#Paper*
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$Ans_{\mathcal{Q}}(AuName, Topic) \longleftarrow Author(AuName, Journal),$
$Journal(Journal, Topic, \#Paper),$

Answers:

$\mathcal{Q}(D)$	*AuName*	*Topic*
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

(John, XML): unexpected answer

What are the causes ?

$\tau =$ *Author(John, TODS)* is an actual cause, with contingency sets

$\Gamma_1=\{$ *Author(John, TKDE)* $\}$ and

$\Gamma_2=\{$*Journal(TKDE, XML, 30)*$\}$

$\rho(\tau) = \frac{1}{2}$

Journal(TKDE, XML, 30), Author(John, TKDE), Journal(TODS,XML, 30) are also actual causes with responsibility $\frac{1}{2}$

Abductive Diagnosis

- Intuitively, an *abductive explanation* of an *observation* is a formula that, together with the *background logical theory*, entails the observation

Abductive Diagnosis

- Intuitively, an *abductive explanation* of an *observation* is a formula that, together with the *background logical theory*, entails the observation
- The background theory specifies the system being diagnosed

Abductive Diagnosis

- Intuitively, an *abductive explanation* of an *observation* is a formula that, together with the *background logical theory*, entails the observation
- The background theory specifies the system being diagnosed
- It is possible to specify systems with *Datalog programs*

Abductive Diagnosis

- Intuitively, an *abductive explanation* of an *observation* is a formula that, together with the *background logical theory*, entails the observation
- The background theory specifies the system being diagnosed
- It is possible to specify systems with *Datalog programs*
  - An example of (recursive) Datalog program:

$$
\begin{array}{rcl}
\multicolumn{3}{l}{parent(mary, john)} \\
\multicolumn{3}{l}{parent(mary, john)} \\
ancestor(X, Y) & \leftarrow & parent(X, Y) \\
ancestor(X, Y) & \leftarrow & parent(X, Z), ancestor(Z, Y)
\end{array}
$$

Abductive Diagnosis

- Intuitively, an *abductive explanation* of an *observation* is a formula that, together with the *background logical theory*, entails the observation
- The background theory specifies the system being diagnosed
- It is possible to specify systems with *Datalog programs*
  - An example of (recursive) Datalog program:

$$\begin{array}{rcl} parent(mary, john) & & \\ parent(mary, john) & & \\ ancestor(X, Y) & \leftarrow & parent(X, Y) \\ ancestor(X, Y) & \leftarrow & parent(X, Z), ancestor(Z, Y) \end{array}$$

- Datalog programs define monotone queries and the notion of cause can be applied as above

Abduction from Datalog Theories

- A *Datalog abduction problem* (DAP) is of the form $\mathcal{AP} = \langle \Pi, E, Hyp, Obs \rangle$,

Abduction from Datalog Theories

- A *Datalog abduction problem* (DAP) is of the form

  $\mathcal{AP} = \langle \Pi, E, Hyp, Obs \rangle$,

  (a) Π: set of Datalog rules

Abduction from Datalog Theories

- A *Datalog abduction problem* (DAP) is of the form

  $\mathcal{AP} = \langle \Pi, E, Hyp, Obs \rangle$,

  (a) Π: set of Datalog rules

  (b) E: finite set of ground atoms (extensional database)

Abduction from Datalog Theories

- A *Datalog abduction problem* (DAP) is of the form

  $\mathcal{AP} = \langle \Pi, E, Hyp, Obs \rangle$,

  (a) Π: set of Datalog rules

  (b) E: finite set of ground atoms (extensional database)

  (c) Hyp: finite set of ground atoms, the abducible atoms

Abduction from Datalog Theories

- A *Datalog abduction problem* (DAP) is of the form

  $\mathcal{AP} = \langle \Pi, E, Hyp, Obs \rangle$,

  (a) Π: set of Datalog rules

  (b) E: finite set of ground atoms (extensional database)

  (c) Hyp: finite set of ground atoms, the abducible atoms

  (d) Obs: observation, a finite conjunction of ground atoms with $\Pi \cup E \cup Hyp \models Obs$

Abduction from Datalog Theories

- A *Datalog abduction problem* (DAP) is of the form

  $\mathcal{AP} = \langle \Pi, E, Hyp, Obs \rangle$,

  (a) Π: set of Datalog rules

  (b) E: finite set of ground atoms (extensional database)

  (c) Hyp: finite set of ground atoms, the abducible atoms

  (d) Obs: observation, a finite conjunction of ground atoms with $\Pi \cup E \cup Hyp \models Obs$

- The *abduction problem* is about computing a subset-minimal $\Delta \subseteq Hyp$, such that $\Pi \cup E \cup \Delta \models Obs$

Abduction from Datalog Theories

- A *Datalog abduction problem* (DAP) is of the form
  $\mathcal{AP} = \langle \Pi, E, Hyp, Obs \rangle$,
  (a) Π: set of Datalog rules
  (b) E: finite set of ground atoms (extensional database)
  (c) Hyp: finite set of ground atoms, the abducible atoms
  (d) Obs: observation, a finite conjunction of ground atoms with $\Pi \cup E \cup Hyp \models Obs$
- The *abduction problem* is about computing a subset-minimal $\Delta \subseteq Hyp$, such that $\Pi \cup E \cup \Delta \models Obs$
- Relevance Problem: Deciding if $h \in Hyp$ belongs to some abductive diagnosis NP-complete! (in $|\mathcal{AP}|$)

View-Update Problem

- View: virtual table defined by a query, e.g. a conjunctive query

View-Update Problem

- View: virtual table defined by a query, e.g. a conjunctive query
- *View-update problem:* update the DB as propagation of changes on views, i.e. from views to base relations

View-Update Problem

- View: virtual table defined by a query, e.g. a conjunctive query
- *View-update problem:* update the DB as propagation of changes on views, i.e. from views to base relations
- When view updates are tuple deletions, this is *delete-propagation* problem

View-Update Problem

- View: virtual table defined by a query, e.g. a conjunctive query
- *View-update problem:* update the DB as propagation of changes on views, i.e. from views to base relations
- When view updates are tuple deletions, this is *delete-propagation* problem
  - How to delete tuples from the database, so that an undesired tuple disappears from the view

Several variants:

View-Update Problem

- View: virtual table defined by a query, e.g. a conjunctive query
- *View-update problem:* update the DB as propagation of changes on views, i.e. from views to base relations
- When view updates are tuple deletions, this is *delete-propagation* problem
  - How to delete tuples from the database, so that an undesired tuple disappears from the view

Several variants:

    - delete subset-minimal set of tuples from source (*minimal source side-effect problem*)

View-Update Problem

- View: virtual table defined by a query, e.g. a conjunctive query
- *View-update problem:* update the DB as propagation of changes on views, i.e. from views to base relations
- When view updates are tuple deletions, this is *delete-propagation* problem
  - How to delete tuples from the database, so that an undesired tuple disappears from the view

Several variants:

- delete subset-minimal set of tuples from source (*minimal source side-effect problem*)
- delete minimum number of tuples from source (*minimum source side-effect problem*)

Delete-Propagation Problem

Author	AuName	Journal
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	Journal	Topic	#Paper
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$$Ans_Q(AuName, Topic) \longleftarrow Author(AuName, Journal), \\ Journal(Journal, Topic, \#Paper),$$

For view V defined by above query:

$V(D)$	$AuName$	$Topic$
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

Don't want (John, XML)
in the view extension

Delete-Propagation Problem

Author	*AuName*	*Journal*
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	*Journal*	*Topic*	*#Paper*
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$$Ans_{\mathcal{Q}}(AuName, Topic) \longleftarrow Author(AuName, Journal), \; Journal(Journal, Topic, \#Paper),$$

For view V defined by above query:

$V(D)$	*AuName*	*Topic*
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

Don't want (John, XML) in the view extension

Minimal source side-effect solutions:

$p_1 = \{$ *Author*(John, TODS), *Journal*(TODS, XML,30)$\}$

$p_2 = \{$ *Author*(John, TODS), *Author*(John, TKDE)$\}$

$p_3 = \{$ *Author*(John, TDK), *Journal*(TODS, XML,30)$\}$

$p_4 = \{$ *Journal*(TODS, XML,30), *Journal*(John, TKDE,30)$\}$

Delete-Propagation Problem

Author	AuName	Journal
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	Journal	Topic	#Paper
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$$Ans_Q(AuName, Topic) \longleftarrow Author(AuName, Journal), Journal(Journal, Topic, \#Paper),$$

For view V defined by above query:

$V(D)$	$AuName$	$Topic$
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

Don't want (John, XML) in the view extension

Minimal source side-effect solutions:

$p_1 = \{$ *Author*(John, TODS), *Journal*(TODS, XML,30)$\}$

$p_2 = \{$ *Author*(John, TODS), *Author*(John, TKDE)$\}$

$p_3 = \{$ *Author*(John, TDK), *Journal*(TODS, XML,30)$\}$

$p_4 = \{$ *Journal*(TODS, XML,30), *Journal*(John, TKDE,30)$\}$

Also solutions to minimum source side-effect problem

Delete-Propagation Problem

Author	*AuName*	*Journal*
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	*Journal*	*Topic*	*#Paper*
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$$Ans_{\mathcal{Q}}(AuName, Topic) \longleftarrow Author(AuName, Journal),$$
$$Journal(Journal, Topic, \#Paper),$$

For view V defined by above query:

$V(D)$	*AuName*	*Topic*
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

Don't want (John, XML) in the view extension

Minimal source side-effect solutions:

p_1={ *Author*(John, TODS), *Journal*(TODS, XML,30)}

p_2={ *Author*(John, TODS), *Author*(John, TKDE)}

p_3={ *Author*(John, TDK), *Journal*(TODS, XML,30)}

p_4={ *Journal*(TODS, XML,30), *Journal*(John, TKDE,30)}

Also solutions to minimum source side-effect problem

Delete-Propagation Problem

Author	AuName	Journal
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	Journal	Topic	#Paper
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$$Ans_{\mathcal{Q}}(AuName, Topic) \longleftarrow Author(AuName, Journal), Journal(Journal, Topic, \#Paper),$$

For view V defined by above query:

$V(D)$	$AuName$	$Topic$
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

Don't want (John, XML) in the view extension

Minimal source side-effect solutions:

$p_1 = \{$ *Author*(John, TODS), *Journal*(TODS, XML,30)$\}$

$p_2 = \{$ *Author*(John, TODS), *Author*(John, TKDE)$\}$

$p_3 = \{$ *Author*(John, TDK), *Journal*(TODS, XML,30)$\}$

$p_4 = \{$ *Journal*(TODS, XML,30), *Journal*(John, TKDE,30)$\}$

Also solutions to minimum source side-effect problem

Notice similarities with causality notions!

Establishing Connections

- In this work we established precise connections between:

Establishing Connections

- In this work we established precise connections between:
  - Causality: a recent problem in databases

Establishing Connections

- In this work we established precise connections between:
  - Causality: a recent problem in databases
  - Datalog abduction

    Abductive diagnosis is an important problem in KR

    Also of interest in databases

Establishing Connections

- In this work we established precise connections between:
  - Causality: a recent problem in databases
  - Datalog abduction

    Abductive diagnosis is an important problem in KR

    Also of interest in databases
  - Deletion-based view updates (interesting case for monotone views)

    View updates is a classical and important problem in databases

Establishing Connections

- In this work we established precise connections between:
  - Causality: a recent problem in databases
  - Datalog abduction

    Abductive diagnosis is an important problem in KR

    Also of interest in databases
  - Deletion-based view updates (interesting case for monotone views)

    View updates is a classical and important problem in databases
- Characterizations of each in terms of the others

  (abduction vs. view-updates has been investigated before)

Establishing Connections

- In this work we established precise connections between:
  - Causality: a recent problem in databases
  - Datalog abduction

    Abductive diagnosis is an important problem in KR

    Also of interest in databases
  - Deletion-based view updates (interesting case for monotone views)

    View updates is a classical and important problem in databases
- Characterizations of each in terms of the others

  (abduction vs. view-updates has been investigated before)
- Known complexity results for some of them are applied to the others, by reduction

Causality and Abduction

- Query-answer causality in databases has a close connection with Datalog abduction

  Causal and abductive entailments coincide

Causality and Abduction

- Query-answer causality in databases has a close connection with Datalog abduction

  Causal and abductive entailments coincide

- Actual causes and their responsibility for a Datalog query can obtained from abductive diagnoses of a corresponding Datalog abduction problem

Causality and Abduction

- Query-answer causality in databases has a close connection with Datalog abduction

  Causal and abductive entailments coincide

- Actual causes and their responsibility for a Datalog query can obtained from abductive diagnoses of a corresponding Datalog abduction problem

- Relevant hypothesis for a Datalog abduction problem can be obtained from actual causes for a corresponding causality problem

Causality and Abduction

Example: Instance D (no exogenous tuples) and Boolean query

$$\Pi: \quad ans \leftarrow R(x,y), S(y),$$

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

Causality and Abduction

Example: Instance D (no exogenous tuples) and Boolean query

$$\Pi: \quad ans \leftarrow R(x,y), S(y),$$

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

Tuples $S(a_1)$, $R(a_2,a_1)$, $S(a_3)$ and $R(a_3,a_3)$ are actual causes for ans, with responsibility $\frac{1}{2}$

Causality and Abduction

Example: (as above) D (no exogenous tuple) Π: $ans \leftarrow R(x,y), S(y)$

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

Causality and Abduction

Example: (as above) D (no exogenous tuple) Π: $ans \leftarrow R(x,y), S(y)$

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

Consider the Datalog abduction problem: $\mathcal{AP}^c = \langle \Pi, \emptyset, D, \{ans\} \rangle$

Causality and Abduction

Example: (as above) D (no exogenous tuple) Π: $ans \leftarrow R(x, y), S(y)$

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

Consider the Datalog abduction problem: $\mathcal{AP}^c = \langle \Pi, \emptyset, D, \{ans\} \rangle$

$\mathcal{AP}^c$ has two (subset-minimal) abductive diagnosis:

Causality and Abduction

Example: (as above) D (no exogenous tuple) $\quad \Pi: \; ans \leftarrow R(x,y), S(y)$

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

Consider the Datalog abduction problem: $\quad \mathcal{AP}^c = \langle \Pi, \emptyset, D, \{ans\} \rangle$

$\mathcal{AP}^c$ has two (subset-minimal) abductive diagnosis:

- $\Delta_1 = \{S(a_1), R(a_2, a_1)\}$

Causality and Abduction

Example: (as above) D (no exogenous tuple) Π: $ans \leftarrow R(x,y), S(y)$

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

Consider the Datalog abduction problem: $\mathcal{AP}^c = \langle \Pi, \emptyset, D, \{ans\} \rangle$

$\mathcal{AP}^c$ has two (subset-minimal) abductive diagnosis:

- $\Delta_1 = \{S(a_1), R(a_2, a_1)\}$
- $\Delta_2 = \{S(a_3), R(a_3, a_3)\}$

Causality and Abduction

Example: (as above) D (no exogenous tuple) $\Pi\colon\ ans \leftarrow R(x,y), S(y)$

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

Consider the Datalog abduction problem: $\mathcal{AP}^c = \langle \Pi, \emptyset, D, \{ans\} \rangle$

$\mathcal{AP}^c$ has two (subset-minimal) abductive diagnosis:

- $\Delta_1 = \{S(a_1), R(a_2, a_1)\}$
- $\Delta_2 = \{S(a_3), R(a_3, a_3)\}$

$S(a_3)$, $R(a_3, a_3)$, $S(a_1)$ and $R(a_2, a_1)$ are relevant hypothesis

Causality and Abduction

Example: (as above) D (no exogenous tuple) $\Pi: \ ans \leftarrow R(x,y), S(y)$

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

Consider the Datalog abduction problem: $\mathcal{AP}^c = \langle \Pi, \emptyset, D, \{ans\} \rangle$

$\mathcal{AP}^c$ has two (subset-minimal) abductive diagnosis:

- $\Delta_1 = \{S(a_1), R(a_2, a_1)\}$
- $\Delta_2 = \{S(a_3), R(a_3, a_3)\}$

$S(a_3)$, $R(a_3, a_3)$, $S(a_1)$ and $R(a_2, a_1)$ are relevant hypothesis

And also the actual causes for ans!

Causality and Abduction

Example: (as above)

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

$\mathcal{AP}^c$ as before, $\Delta_1 = \{S(a_1), R(a_2, a_1)\}$, $\Delta_2 = \{S(a_3), R(a_3, a_3)\}$

Causality and Abduction

Example: (as above)

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

$\mathcal{AP}^c$ as before, $\Delta_1 = \{S(a_1), R(a_2, a_1)\}$, $\Delta_2 = \{S(a_3), R(a_3, a_3)\}$

To obtain responsibilities, we compute *necessary-hypothesis sets* of $\mathcal{AP}^c$

Causality and Abduction

Example: (as above)

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

$\mathcal{AP}^c$ as before, $\Delta_1 = \{S(a_1), R(a_2, a_1)\}$, $\Delta_2 = \{S(a_3), R(a_3, a_3)\}$

To obtain responsibilities, we compute *necessary-hypothesis sets* of $\mathcal{AP}^c$

W/O them in *Hyp* there is no abductive diagnosis

Causality and Abduction

Example: (as above)

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

$\mathcal{AP}^c$ as before, $\Delta_1 = \{S(a_1), R(a_2, a_1)\}$, $\Delta_2 = \{S(a_3), R(a_3, a_3)\}$

To obtain responsibilities, we compute *necessary-hypothesis sets* of $\mathcal{AP}^c$

W/O them in *Hyp* there is no abductive diagnosis

- $N_1 = \{S(a_1), S(a_3)\}$, $N_2 = \{S(a_3), R(a_2, a_1)\}$
- $N_3 = \{S(a_1), R(a_3, a_3)\}$, $N_4 = \{R(a_2, a_1), R(a_3, a_3)\}$

Causality and Abduction

Example: (as above)

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

$\mathcal{AP}^c$ as before, $\Delta_1 = \{S(a_1), R(a_2, a_1)\}$, $\Delta_2 = \{S(a_3), R(a_3, a_3)\}$

To obtain responsibilities, we compute *necessary-hypothesis sets* of $\mathcal{AP}^c$

W/O them in *Hyp* there is no abductive diagnosis

- $N_1 = \{S(a_1), S(a_3)\}$, $N_2 = \{S(a_3), R(a_2, a_1)\}$
- $N_3 = \{S(a_1), R(a_3, a_3)\}$, $N_4 = \{R(a_2, a_1), R(a_3, a_3)\}$

It holds: The responsibilities of causal tuples τ for *ans* are $\frac{1}{|N|}$, with N a minimum-cardinality necessary-hypothesis set, and $\tau \in N$

Causality and Abduction

Example: (as above)

R	X	Y
	a_1	a_4
	a_2	a_1
	a_3	a_3

S	X
	a_1
	a_2
	a_3

$\mathcal{AP}^c$ as before, $\Delta_1 = \{S(a_1), R(a_2, a_1)\}$, $\Delta_2 = \{S(a_3), R(a_3, a_3)\}$

To obtain responsibilities, we compute *necessary-hypothesis sets* of $\mathcal{AP}^c$

W/O them in *Hyp* there is no abductive diagnosis

- $N_1 = \{S(a_1), S(a_3)\}$, $N_2 = \{S(a_3), R(a_2, a_1)\}$
- $N_3 = \{S(a_1), R(a_3, a_3)\}$, $N_4 = \{R(a_2, a_1), R(a_3, a_3)\}$

It holds: The responsibilities of causal tuples τ for *ans* are $\frac{1}{|N|}$, with N a minimum-cardinality necessary-hypothesis set, and $\tau \in N$

Again, all causes have responsibility $\frac{1}{2}$

Causality and Abduction

- These connections enable mutual applications

Causality and Abduction

- These connections enable mutual applications
- For example, for Datalog queries, deciding if a tuple is a cause is *NP*-complete (combined complexity)

Causality and Abduction

- These connections enable mutual applications
- For example, for Datalog queries, deciding if a tuple is a cause is *NP*-complete (combined complexity)
  - Obtained from complexity of relevance problem for Datalog abduction

Causality and Abduction

- These connections enable mutual applications
- For example, for Datalog queries, deciding if a tuple is a cause is *NP*-complete (combined complexity)
  - Obtained from complexity of relevance problem for Datalog abduction
  - Notice that for unions of Boolean conjunctive queries the problem is tractable in data (Salimi & Bertossi, ICDT 2015)

Causality and Abduction

- These connections enable mutual applications
- For example, for Datalog queries, deciding if a tuple is a cause is *NP*-complete (combined complexity)
  - Obtained from complexity of relevance problem for Datalog abduction
  - Notice that for unions of Boolean conjunctive queries the problem is tractable in data (Salimi & Bertossi, ICDT 2015)
- We identify tractable classes of deciding causality for Datalog queries and instances

Causality and Abduction

- These connections enable mutual applications
- For example, for Datalog queries, deciding if a tuple is a cause is *NP*-complete (combined complexity)
  - Obtained from complexity of relevance problem for Datalog abduction
  - Notice that for unions of Boolean conjunctive queries the problem is tractable in data (Salimi & Bertossi, ICDT 2015)
- We identify tractable classes of deciding causality for Datalog queries and instances
  - Obtained from tractable cases of Datalog abduction

    Guarded programs, bounded tree-width instances

    (Gottlob, Pichler & Wei, 2010)

Causality and Delete Propagation

- As suggested above, there is a close relationship between query causality and delete-propagation for view-updates

Causality and Delete Propagation

- As suggested above, there is a close relationship between query causality and delete-propagation for view-updates
- Want to propagate deletion a tuple $\bar{v}$ from view V to underlying D?

Causality and Delete Propagation

- As suggested above, there is a close relationship between query causality and delete-propagation for view-updates
- Want to propagate deletion a tuple $\bar{v}$ from view V to underlying D?
- If V defined by monotone query $\mathcal{Q}(\bar{x})$, delete from D the actual causes for answer $\bar{v}$

Causality and Delete Propagation

- As suggested above, there is a close relationship between query causality and delete-propagation for view-updates
- Want to propagate deletion a tuple $\bar{v}$ from view V to underlying D?
- If V defined by monotone query $\mathcal{Q}(\bar{x})$, delete from D the actual causes for answer $\bar{v}$
- Actual causes with subset-minimal contingency sets give solutions to minimal source side-effect problem (and viceversa)

Causality and Delete Propagation

- As suggested above, there is a close relationship between query causality and delete-propagation for view-updates
- Want to propagate deletion a tuple $\bar{v}$ from view V to underlying D?
- If V defined by monotone query $\mathcal{Q}(\bar{x})$, delete from D the actual causes for answer $\bar{v}$
- Actual causes with subset-minimal contingency sets give solutions to minimal source side-effect problem (and viceversa)
- Most responsible actual causes with minimum-cardinality contingency sets give solutions to minimum source side-effect problem (and viceversa)

Causality and Delete Propagation

Example: Again

Author	AuName	Journal
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	Journal	Topic	#Paper
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$$Ans_{\mathcal{Q}}(AuName, Topic) \longleftarrow Author(AuName, Journal), Journal(Journal, Topic, \#Paper),$$

View:

$V(D)$	AuName	Topic
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

(John, XML) not wanted in the view

Causality and Delete Propagation

Example: Again

Author	AuName	Journal
	Joe	TKDE
	John	TKDE
	Tom	TKDE
	John	TODS

Journal	Journal	Topic	#Paper
	TKDE	XML	30
	TKDE	CUBE	30
	TODS	XML	30

$$Ans_{\mathcal{Q}}(AuName, Topic) \longleftarrow Author(AuName, Journal), Journal(Journal, Topic, \#Paper),$$

View:

V(D)	AuName	Topic
	Joe	XML
	Joe	CUBE
	Tom	XML
	Tom	CUBE
	John	XML
	John	CUBE

(John, XML) not wanted in the view

Combination for (John, XML) of actual causes/contingency sets

(each element is actual cause and the complement a contingency set)

p_1={ *Author*(John, TODS), *Journal*(TODS, XML,30)}

p_2={ *Author*(John, TODS), *Author*(John, TKDE)}

p_3={ *Author*(John, TDK), *Journal*(TODS, XML,30)}

p_4={ *Journal*(TODS, XML,30), *Journal*(John, TKDE,30)}

Causality and Delete Propagation

We can take advantage of established connections:

Causality and Delete Propagation

We can take advantage of established connections:

- Computing the size of a solution to a minimum source side-effect problem is $FP^{NP(log(n))}$-hard in data

  From complexity of computing most responsible causes in

  (Salimi and Bertossi ICDT, 2015)

- We identify class of queries for which the minimum source side-effect problem is tractable

  From the dichotomy result for complexity of responsibility in

  (Meliou et al., VLDB 2010)

Conclusions

- Causality has been a research subject in AI, Statistics, etc. etc. for many years

Conclusions

- Causality has been a research subject in AI, Statistics, etc. etc. for many years
- Causality in data management (DM) is much newer subject

Conclusions

- Causality has been a research subject in AI, Statistics, etc. etc. for many years
- Causality in data management (DM) is much newer subject
- We have started to scratch the surface (as a DM community)

Conclusions

- Causality has been a research subject in AI, Statistics, etc. etc. for many years
- Causality in data management (DM) is much newer subject
- We have started to scratch the surface (as a DM community)
- Many extensions are possible and necessary, to understand data phenomena in causal terms

Conclusions

- Causality has been a research subject in AI, Statistics, etc. etc. for many years
- Causality in data management (DM) is much newer subject
- We have started to scratch the surface (as a DM community)
- Many extensions are possible and necessary, to understand data phenomena in causal terms
- Interesting results on causality in DM have been obtained

Conclusions

- Causality has been a research subject in AI, Statistics, etc. etc. for many years
- Causality in data management (DM) is much newer subject
- We have started to scratch the surface (as a DM community)
- Many extensions are possible and necessary, to understand data phenomena in causal terms
- Interesting results on causality in DM have been obtained
- Causality in DM is related to many other DM reasoning tasks

Conclusions

- Causality has been a research subject in AI, Statistics, etc. etc. for many years
- Causality in data management (DM) is much newer subject
- We have started to scratch the surface (as a DM community)
- Many extensions are possible and necessary, to understand data phenomena in causal terms
- Interesting results on causality in DM have been obtained
- Causality in DM is related to many other DM reasoning tasks
- Maybe it is -unsurprisingly- an important underlying principle

Conclusions

- Causality has been a research subject in AI, Statistics, etc. etc. for many years
- Causality in data management (DM) is much newer subject
- We have started to scratch the surface (as a DM community)
- Many extensions are possible and necessary, to understand data phenomena in causal terms
- Interesting results on causality in DM have been obtained
- Causality in DM is related to many other DM reasoning tasks
- Maybe it is -unsurprisingly- an important underlying principle
- Or one that can lead us to a unifying concept other tasks may emerge from