On the Role of the Complementation Rule for Data Dependencies over Incomplete Relations

Sven Hartmann

Department of Informatics, Clausthal University of Technology, Germany

Flavio Ferrarotti, Sebastian Link

School of Information Management, Victoria University of Wellington, New Zealand

UnB

MARSDEN FUND
TE PŪTEA RANGAHAU A MARSDEN

Main contributions

- investigate inference systems for a class of data dependencies
  - ↪ functional & multivalued dependencies, & NOT NULL constraints
  - ↪ *complementation rule* may infer meaningless dependencies
  - ↪ Can axiomatizations depict necessarily meaningful dependencies?
- recently proposed axiomatization is inappropriate
- addition of further inference rules leads to an appropriate system
- indicates that complementation rule is mere means of normalization

UnB

MARSDEN FUND
TE PŪTEA RANGAHAU A MARSDEN

Running example - Informal

- suppliers deliver articles from a location at a certain cost:

```
CREATE TABLE SUPPLIES (
    Article CHAR[20],
    Supplier VARCHAR NOT NULL,
    Location VARCHAR NOT NULL,
    Cost CHAR[8]);
```

- DBMS explicitly enforces additional business rules:
  - ↪ for every article there is at most one supplier,
  - ↪ cost determined by article and location, and
  - ↪ location sets determined by supplier (independent of article & cost)
- are the following business rules already enforced implicitly?
  - ↪ cost is determined by article
  - ↪ location sets determined by article (independent of supplier & cost)

Running Example - Formal

- relation schema R:

$$\textsc{Supplies}=\{\text{A(rticle), S(upplier), L(ocation), C(ost)}\}$$

- *no-information* null value $\mathtt{ni}$ (distinguished value of each domain)
  $\hookrightarrow$ value may not exist, or value exists and is unknown [Zaniolo]

- null-free subschema NFS $R_s \subseteq R$ over R [Atzeni/Morfuni]:

$$\textsc{Supplies}_S=\{\text{Supplier, Location}\}$$

  $\hookrightarrow$ r satisfies R_s iff every $t \in r$ is R_s-total $(\forall A \in R_s(t(A) \neq \mathtt{ni}))$

- data dependencies enforce business rules:
  $\hookrightarrow$ FDs: $A \to S$ and $AL \to C$
  $\hookrightarrow$ MVD: $S \twoheadrightarrow L$

UnB

MARSDEN FUND
TE PŪTEA RANGAHAU A MARSDEN

Data Dependencies in Presence of No-information Nulls

- relation r satisfies FD $X \to Y$ over R [Lien, Atzeni/Morfuni]:
  $\hookrightarrow$ if for all tuples $t_1, t_2 \in r$ we have:

  if $t_1(X) = t_2(X)$ and t_1, t_2 are X-total, then $t_1(Y) = t_2(Y)$

- relation r satisfies MVD $X \twoheadrightarrow Y$ over R [Lien]:
  $\hookrightarrow$ if for all tuples $t_1, t_2 \in r$ we have:

  if $t_1(X) = t_2(X)$ and t_1, t_2 are X-total, then there is some tuple t in r such that $t(XY) = t_1(XY)$ and $t(X(R-Y)) = t_2(X(R-Y))$

- relationship to decompositions [Lien]:
  $\hookrightarrow$ if r satisfies $X \to Y$, then $r_X[R] = r_X[XY] \bowtie r_X[X(R-Y)]$
  $\hookrightarrow$ r satisfies $X \twoheadrightarrow Y$ if and only if $r_X[R] = r_X[XY] \bowtie r_X[X(R-Y)]$

A Decomposition Example

- $R = ASLC$, $R_s = ALC$, $\Sigma = \{A \rightarrow S; AL \rightarrow C; S \twoheadrightarrow L\}$
  $\hookrightarrow$ decompose into $R_1 = SL$ and $R_2 = ASC$

r			
Article	Supplier	Location	Cost
Kiwi	ni	Maunganui	1.50
Kiwi	ni	Taranaki	2.50
Green Kiwi	G6Kiwi	Wellington	1.50
Gold Kiwi	G6Kiwi	Christchurch	2.50
Green Kiwi	G6Kiwi	Christchurch	1.50
Gold Kiwi	G6Kiwi	Wellington	2.50

$r[SL]$	
Supplier	Location
ni	Maunganui
ni	Taranaki
G6Kiwi	Wellington
G6Kiwi	Christchurch

$r[ASC]$		
Article	Supplier	Cost
Kiwi	ni	1.50
Kiwi	ni	2.50
Green Kiwi	G6Kiwi	1.50
Gold Kiwi	G6Kiwi	2.50

- $r_S[R] = r_S[SL] \bowtie r_S[ASC]$:

$r_S[R]$			
Article	Supplier	Location	Cost
Green Kiwi	G6Kiwi	Wellington	1.50
Gold Kiwi	G6Kiwi	Christchurch	2.50
Green Kiwi	G6Kiwi	Christchurch	1.50
Gold Kiwi	G6Kiwi	Wellington	2.50

$r_S[SL]$	
Supplier	Location
G6Kiwi	Wellington
G6Kiwi	Christchurch

$r_S[ASC]$		
Article	Supplier	Cost
Green Kiwi	G6Kiwi	1.50
Gold Kiwi	G6Kiwi	2.50

The R-Implication Problem

- implication problem:
  $\hookrightarrow$ input: R, NFS R_s and FD/MVD set $\Sigma \cup \{\varphi\}$ over R
  $\hookrightarrow$ output: $\begin{cases} \text{yes}, \text{ if } \Sigma \models_{R_s} \varphi \\ \text{no}, \text{ if } \Sigma \not\models_{R_s} \varphi \end{cases}$ where $\Sigma \models_{R_s} \varphi$ holds, if every relation over R that satisfies Σ and R_s also satisfies φ

- $R = ASLC$, $R_s = SL$, $\Sigma = \{A \to S; AL \to C; S \twoheadrightarrow L\}$:
  $\hookrightarrow$ $\Sigma \models_{R_s} A \to C$ and $\Sigma \models_{R_s} A \twoheadrightarrow L$

- if $R_s = ALC$ instead, then $\Sigma \not\models_{R_s} A \to C$ and $\Sigma \not\models_{R_s} A \twoheadrightarrow L$:

Article	Supplier	Location	Cost
Kiwi	`ni`	Maunganui	1.50
Kiwi	`ni`	Taranaki	2.50

Outline

- Current Axiomatization and Inferences
- Desiderata for an Appropriate Axiomatization
  $\hookrightarrow$ Inappropriateness of Current Axiomatization
- An Appropriate Axiomatization and Inferences
- Implication in Undetermined Universes
- Related Work
- Conclusion and Open Problems

The FD/MVD Axiomatization $\mathfrak{D}$ for arbitrary NFS R_s

$$\frac{}{XY \to Y}$$
(reflexivity, $\mathcal{R}_{\text{F}}$)

$$\frac{X \to YZ}{X \to Y}$$
(decomposition, $\mathcal{D}_{\text{F}}$)

$$\frac{X \to Y; X \to Z}{X \to YZ}$$
(FD union, $\mathcal{U}_{\text{F}}$)

$$\frac{X \twoheadrightarrow Y}{X \twoheadrightarrow R - Y}$$
(R-complementation, $\mathcal{C}^R_{\text{M}}$)

$$\frac{X \twoheadrightarrow Y; X \twoheadrightarrow Z}{X \twoheadrightarrow YZ}$$
(MVD union, $\mathcal{U}_{\text{M}}$)

$$\frac{X \twoheadrightarrow W; Y \twoheadrightarrow Z}{X \twoheadrightarrow Z - W} \, Y \subseteq X(W \cap R_s)$$
(null pseudo-transitivity, $\mathcal{T}_{\text{M}}$)

$$\frac{X \to Y}{X \twoheadrightarrow Y}$$
(implication, $\mathcal{I}_{\text{FM}}$)

$$\frac{X \twoheadrightarrow W; Y \to Z}{X \to Z - W} \, Y \subseteq X(W \cap R_s)$$
(null mixed pseudo-transitivity, $\mathcal{T}_{\text{FM}}$)

$\hookrightarrow$ S. Hartmann, S. Link: When data dependencies over SQL tables meet the Logics of Paradox and S-3, 29th ACM Symposium on Principles of Database Systems 2010, p:317-326.

UnB

MARSDEN FUND
TE PŪTEA RANGAHAU A MARSDEN

Some Inferences in $\mathfrak{D}$

- $R = ASLC$, $R_s = SL$, $\Sigma = \{A \rightarrow S; AL \rightarrow C; S \twoheadrightarrow L\}$

$$\dfrac{\dfrac{\dfrac{\dfrac{A \rightarrow S}{\mathcal{I}_{\mathrm{FM}} : A \twoheadrightarrow S} \quad \dfrac{S \twoheadrightarrow AC}{\mathcal{C}^{R}_{\mathrm{M}} : S \twoheadrightarrow SL}}{\mathcal{T}_{\mathrm{M}} : A \twoheadrightarrow L} \quad \dfrac{\dfrac{}{\mathcal{R}_{\mathrm{F}} : A \rightarrow A}}{\mathcal{I}_{\mathrm{FM}} : A \twoheadrightarrow A}}{\mathcal{U}_{\mathrm{M}} : A \twoheadrightarrow AL} \quad AL \rightarrow C}{\mathcal{T}_{\mathrm{FM}} : A \rightarrow C}$$

- Can we infer $A \twoheadrightarrow L$ with either no application of $\mathcal{C}^{R}_{\mathrm{M}}$ or just an application of $\mathcal{C}^{R}_{\mathrm{M}}$ in the very last step of the inference?
- Can we infer $A \rightarrow C$ without an application of $\mathcal{C}^{R}_{\mathrm{M}}$?
- Are $A \twoheadrightarrow L$ and $A \rightarrow C$ semantically meaningful constraints?

Desiderata for Inference Systems

- $\mathfrak{R}$ a set of inference rules for R-implication of FDs, MVDs and an NFS, R-complementation rule $\mathcal{C}^R_{\mathrm{M}}$ only rule that is dependent on R
- $\mathfrak{R}$ *complementary* for R-implication of FDs and MVDs, if
  - $\hookrightarrow$ for every R, for every NFS R_s, every FD/MVD set Σ over R, and
  - $\hookrightarrow$ every MVD φ over R that is R-implied by Σ in the presence of R_s,
  - $\hookrightarrow$ there is an inference of φ from Σ by $\mathfrak{R}$ in which
  - $\hookrightarrow$ $\mathcal{C}^R_{\mathrm{M}}$ applied at most once and if, then only in last step
- $\mathfrak{R}$ *adequate* for R-implication of FDs and MVDs, if
  - $\hookrightarrow$ for every R, for every NFS R_s, every FD/MVD set Σ over R, and
  - $\hookrightarrow$ every FD φ over R that is R-implied by Σ in the presence of R_s,
  - $\hookrightarrow$ there is an inference of φ from Σ by $\mathfrak{R}$ in which
  - $\hookrightarrow$ $\mathcal{C}^R_{\mathrm{M}}$ is not applied at all
- $\mathfrak{R}$ *appropriate* if $\mathfrak{R}$ is complementary and adequate.

$\mathfrak{D}$ is neither complementary nor adequate

- $\mathfrak{D}$ is not complementary:
  - $\hookrightarrow R = ABCD = R_s$, $\Sigma = \{\emptyset \twoheadrightarrow AB; C \twoheadrightarrow A\}$, and $\varphi = \emptyset \twoheadrightarrow A$
  - $\hookrightarrow \varphi \notin \Sigma^+_{\mathfrak{D}-\{\mathcal{C}^R_M\}}$
  - $\hookrightarrow \varphi \in \Sigma^+_{\mathfrak{D}}$: $\mathcal{C}^R_M$ must be applied to infer φ from Σ
  - $\hookrightarrow$ since $\emptyset \twoheadrightarrow BCD \notin \Sigma^+_{\mathfrak{D}-\{\mathcal{C}^R_M\}}$, $\mathcal{C}^R_M$ not just applied in the last step
- $\mathfrak{D}$ is not adequate:
  - $\hookrightarrow R = AB = R_s$, $\Sigma = \{\emptyset \twoheadrightarrow A; B \rightarrow A\}$, and $\varphi = \emptyset \rightarrow A$
  - $\hookrightarrow \varphi \notin \Sigma^+_{\mathfrak{D}-\{\mathcal{C}^R_M\}}$, but $\varphi \in \Sigma^+_{\mathfrak{D}}$

UnB MARSDEN FUND TE PŪTEA RANGAHAU A MARSDEN

The Inference System $\mathfrak{F}$

$$\frac{}{XY \rightarrow Y}$$
(reflexivity, $\mathcal{R}_{\mathrm{F}}$)

$$\frac{X \rightarrow YZ}{X \rightarrow Y}$$
(decomposition, $\mathcal{D}_{\mathrm{F}}$)

$$\frac{X \rightarrow Y; X \rightarrow Z}{X \rightarrow YZ}$$
(FD union, $\mathcal{U}_{\mathrm{F}}$)

$$\frac{X \twoheadrightarrow Y; X \twoheadrightarrow Z}{X \twoheadrightarrow YZ}$$
(MVD union, $\mathcal{U}_{\mathrm{M}}$)

$$\frac{X \twoheadrightarrow W; Y \twoheadrightarrow Z}{X \twoheadrightarrow Z - W} \; Y \subseteq X(W \cap R_s)$$
(null pseudo-transitivity, $\mathcal{T}_{\mathrm{M}}$)

$$\frac{X \twoheadrightarrow Y}{X \twoheadrightarrow R - Y}$$
(R-complementation, $\mathcal{C}_{\mathrm{M}}^{R}$)

$$\frac{X \twoheadrightarrow W; Y \twoheadrightarrow Z}{X \twoheadrightarrow ZW} \; Y \subseteq X(W \cap R_s)$$
(**additive null transitivity**, $\mathcal{T}_{\mathrm{M}}^{*}$)

$$\frac{X \twoheadrightarrow W; Y \twoheadrightarrow Z}{X \twoheadrightarrow Z \cap W} \; Y \subseteq XR_s; \; (Y - X) \cap W = \emptyset$$
(**null subset**, $\mathcal{S}_{\mathrm{M}}$)

$$\frac{X \rightarrow Y}{X \twoheadrightarrow Y}$$
(implication, $\mathcal{I}_{\mathrm{FM}}$)

$$\frac{X \twoheadrightarrow W; Y \rightarrow Z}{X \rightarrow Z \cap W} \; Y \subseteq XR_s; \; (Y - X) \cap W = \emptyset$$
(**null mixed subset**, $\mathcal{S}_{\mathrm{FM}}$)

$$\frac{X \twoheadrightarrow W; Y \rightarrow Z}{X \rightarrow Z - W} \; Y \subseteq X(W \cap R_s)$$
(null mixed pseudo-transitivity, $\mathcal{T}_{\mathrm{FM}}$)

$\mathfrak{F}$ is indeed appropriate

- *Theorem:* Let Σ be a set of FDs and MVDs over relation schema R, and $R_s \subseteq R$. For every inference γ from Σ by the system $\mathfrak{D}$ there is an inference ξ from Σ by the system $\mathfrak{F}$ with the following properties:

(1) if γ infers an MVD, then

- γ and ξ infer the same MVD,
- in ξ the R-complementation rule is applied at most once, and
- if the R-complementation rule is applied in ξ, then it is applied as the last rule.

(2) if γ infers an FD, then

- γ and ξ infer the same FD, and
- in ξ the R-complementation rule is not applied at all. □

Some appropriate Inferences

- $R = ASLC$, $R_s = SL$, $\Sigma = \{A \rightarrow S; AL \rightarrow C; S \twoheadrightarrow AC\}$

$\hookrightarrow$ infer $A \twoheadrightarrow L$ by applying $\mathcal{C}_{\mathrm{M}}^{R}$ in last step only

$$\cfrac{\cfrac{\mathcal{I}_{\mathrm{FM}}: \cfrac{A \rightarrow S}{A \twoheadrightarrow S} \quad S \twoheadrightarrow AC}{\mathcal{T}_{\mathrm{M}}^{*}: \quad A \twoheadrightarrow ACS}}{\mathcal{C}_{\mathrm{M}}^{R}: \quad A \twoheadrightarrow L}$$

$\hookrightarrow$ infer $A \rightarrow C$ without an application of $\mathcal{C}_{\mathrm{M}}^{R}$

$$\cfrac{\cfrac{\mathcal{I}_{\mathrm{FM}}: \cfrac{A \rightarrow S}{A \twoheadrightarrow S} \quad S \twoheadrightarrow AC}{\mathcal{T}_{\mathrm{M}}^{*}: \quad A \twoheadrightarrow ACS} \quad AL \rightarrow C}{\mathcal{S}_{\mathrm{FM}}: \quad A \rightarrow C}$$

$\hookrightarrow$ note that $AL \subseteq ALS$ and $(AL - A) \cap ACS = \emptyset$

$\mathfrak{U}$ is almost R-complete

- Let $\mathfrak{U} = \mathfrak{F} - \{\mathcal{C}^R_{\mathrm{M}}\}$
- *Corollary:* Let $\Sigma \cup \{\varphi\}$ be a finite set of FDs and MVDs over the relation schema R. Then
  - If φ denotes an FD, then: $\varphi \in \Sigma^+_{\mathfrak{F}}$ if and only if $\varphi \in \Sigma^+_{\mathfrak{U}}$.
  - If φ denotes the MVD $X \twoheadrightarrow Y$, then: $X \twoheadrightarrow Y \in \Sigma^+_{\mathfrak{F}}$ if and only if $X \twoheadrightarrow Y \in \Sigma^+_{\mathfrak{U}}$ or $X \twoheadrightarrow (R - Y) \in \Sigma^+_{\mathfrak{U}}$. □

Implication over undetermined Universes

- $Dom(r)$ denotes the set of attributes over which r is defined
- $Attr(X \twoheadrightarrow Y) = Attr(X \rightarrow Y) = XY$
- Let $\Sigma \cup \{\varphi\}$ be a set of FDs and MVDs and R_s an NFS. We say that Σ *implies* φ in the presence of R_s if and only if every relation r satisfies the following condition: if $\bigcup_{\sigma \in \Sigma} Attr(\sigma) \cup Attr(\varphi) \cup R_s \subseteq Dom(r)$ and r satisfies all $\sigma \in \Sigma$ and R_s, then r also satisfies φ.

$\mathfrak{U}$ is an axiomatization over undetermined Universes

- R-implication does not imply implication:
  - $\hookrightarrow \Sigma = \{A \to S, AL \to C, S \twoheadrightarrow AC\}$ and $R_s = SL$
  - $\hookrightarrow A \twoheadrightarrow L$ is $ASCL$-implied by Σ and R_s
  - $\hookrightarrow A \twoheadrightarrow L$ is not implied by Σ and R_s:

Supplier	Article	Cost	Location	Quantity
Taratua&Co	Kea	`ni`	Gisborne	2
Taratua&Co	Kea	`ni`	Wellington	3

- *Theorem:* The set $\mathfrak{U}$ is a finite axiomatization for the implication of FDs and MVDs in the presence of an NFS over undetermined universes. □
  - $\hookrightarrow A \twoheadrightarrow L$ is not necessarily a semantically meaningful constraint
  - $\hookrightarrow A \to C$ is necessarily a semantically meaningful constraint

Previous Work on this Topic

- over total relations (every attribute is NOT NULL)
  - ↪ Biskup: complementarity for MVDs
  - ↪ Biskup/Link: added adequacy to combined FD/MVD set
- over partial relations (no attribute is NOT NULL)
  - ↪ Link: complementarity for MVDs
  - ↪ in this case MVDs and FDs interact trivially
- our theory unifies these two works
- other work:
  - ↪ Hartmann/Link: complementarity for weak MVDs
  - ↪ Hartmann/Link: FDs and full hierarchical dependencies
  - ↪ Link: full first-order hierarchical decompositions

Conclusion

- investigated inferences for a class of database dependencies
  - ↪ functional & multivalued dependencies, & NOT NULL constraints
  - ↪ normalization dictates restricted apps of complementation rule
  - ↪ studied if all inferences implement restricted apps appropriately
- recently proposed axiomatization is inappropriate
- addition of further inference rules leads to an appropriate system
- studied notion of implication where underlying universe undetermined
  - ↪ axiomatization does not require the complementation rule
- indicates that complementation rule is mere means of normalization

Future Work

- investigate other null values, at least:
  - ↪ not applicable
  - ↪ does not exist
  - ↪ exists, but unknown (theory valid under possible world semantics)
- probability theory: correspondences to conditional independencies
- study FDs, full hierarchical dependencies and NOT NULL constraints
- investigate these dependencies also over OR-relations