

# Normalisation in the Presence of Lists

Sven Hartmann, Sebastian Link

Information Science Research Centre,  
Massey University, Palmerston North,  
New Zealand

1. Motivation & Revision of the RDM
2. The Brouwerian Algebra of Nested Attributes
3. Axiomatisation of FDs
4. Redundancies and the Nested List Normal Form
5. Update Anomalies
6. Future Work

## 1.1 A first Example

- consider **Factor(Integer,Prime[Number],Exp[Number])**
- a typical snapshot  $r$  consists of the following tuples

(12,[2,3],[2,1]),  
(35,[5,7],[1,1]),  
(37,[37],[1]),  
(936,[2,3,13],[3,2,1])

- FDs on **Factor(Integer,Prime[Number],Exp[Number])** are
  - Factor(Integer), Factor(Prime[Number],Exp[Number]) minimal keys
  - Factor(Prime[ $\lambda$ ])  $\rightarrow$  Factor(Exp[ $\lambda$ ])
  - Factor(Exp[ $\lambda$ ])  $\rightarrow$  Factor(Prime[ $\lambda$ ])
- Is this a well-designed database? Why? Why not?

## 1.2 Boyce-Codd Normal Form in the RDM

- **FDs** with  $X, Y \subseteq R$  introduced by E.F. Codd in 1972
  - $\models_r X \rightarrow Y$  iff  $t_1[Y] = t_2[Y]$ , if  $t_1[X] = t_2[X]$  for any  $t_1, t_2 \in r$

- **Armstrong Axioms**

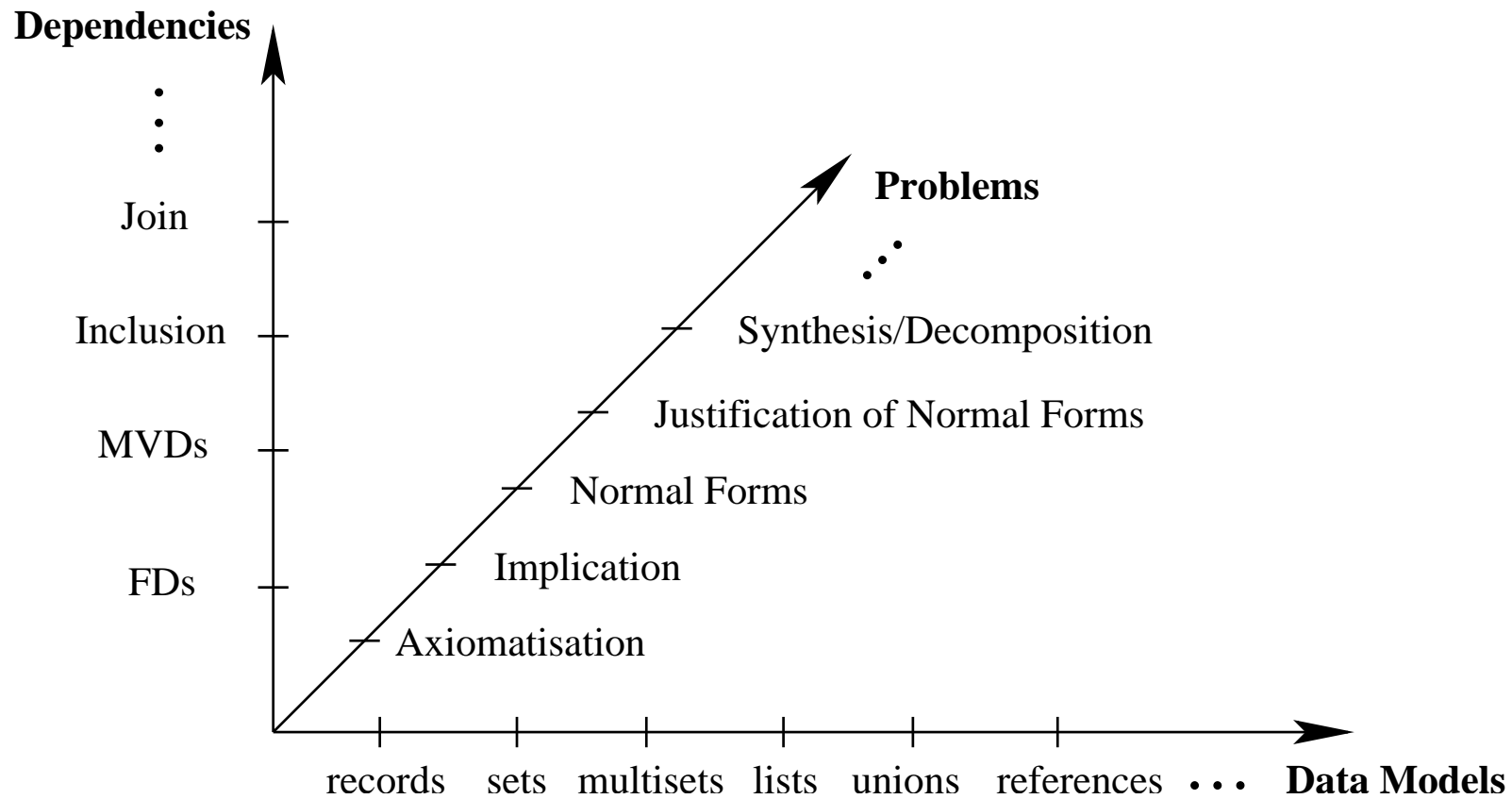
$$\frac{}{X \rightarrow Y} Y \subseteq X \qquad \frac{X \rightarrow Y}{X \rightarrow X \cup Y} \qquad \frac{X \rightarrow Y, Y \rightarrow Z}{X \rightarrow Z}$$

form minimal, sound and complete set of Inference Rules

- $(R, \Sigma)$  in **BCNF** iff every  $X \rightarrow Y \in \Sigma^*$  trivial or  $X$  is  $R$ -superkey
- justification:  $(R, \Sigma)$  in BCNF
  - iff every  $X \rightarrow Y \in \Sigma$  is trivial or  $X$  superkey for  $R$
  - iff every  $R$ -relation  $r$  with  $\models_r \Sigma_{\text{key}}$  satisfies  $\models_r \Sigma$
  - iff  $R$  non-redundant wrt.  $\Sigma$
  - iff  $R$  does not have any insertion anomalies
  - iff  $R$  does not have any replacement anomalies (type 1 or 2)
  - only if  $R$  does not have any type 3 replacement anomalies

## 1.3 Challenges with Advanced Data Models

- find unifying framework, extend achievements to complex object types
- classify data models according to the **types** they support



## 1.4 The Need for Lists

- here: **values, records of values and lists of values**
- ordered relations
- time-series data, meteorological and astronomical data streams, runs of experimental data, multidimensional arrays, textual information, voices, sound, images, video, etc.
- subject to studies in deductive and temporal database community
- occur naturally in object-oriented databases (XML)
- bioinformatics: lists occur naturally in genomic sequence databases

## 2.1 Syntax: Nested Attributes

- capture characteristics of objects in target database by attributes
- finite set  $\mathcal{U}$  of flat attributes and  $dom(A)$  for all  $A \in \mathcal{U}$
- use set  $\mathcal{L}$  of labels with  $\mathcal{U} \cap \mathcal{L} = \emptyset$  and  $\lambda \notin \mathcal{U} \cup \mathcal{L}$
- **nested attributes**  $\mathcal{NA}(\mathcal{U}, \mathcal{L})$ :
  - *flat attributes*  $\mathcal{U} \subseteq \mathcal{NA}$ ,
  - *null attribute*  $\lambda \in \mathcal{NA}$ ,
  - *record-valued attributes*  $L(N_1, \dots, N_k) \in \mathcal{NA}$ , if  $L \in \mathcal{L}$  and  $N_1, \dots, N_k \in \mathcal{NA}$  with  $k \geq 1$
  - *list-valued attributes*  $L[N] \in \mathcal{NA}$ , if  $L \in \mathcal{L}$  and  $N \in \mathcal{NA}$
- Factor(Integer, Prime[Number], Exp[Number]) is nested attribute

## 2.2 Semantics: Domain Assignment

- extend mapping  $dom$  from flat attributes to nested attributes by:
  - $dom(\lambda) = \{ok\}$ ,
  - $dom(L(N_1, \dots, N_k)) = \{(v_1, \dots, v_k) \mid v_i \in dom(N_i)\}$ ,
  - $dom(L[N]) = \{[v_1, \dots, v_n] \mid v_i \in dom(N)\}$
- empty list denoted by  $[]$
- RDM: record-valued attributes only
- $Dom(\text{Factor}(\text{Integer}, \text{Prime}[\text{Number}], \text{Exp}[\text{Number}]))$ : 3-tuples where
  - first component is some integer value and
  - second and third components are both a finite list of integer values

## 2.3 Subattributes

- define  $\leq \subseteq \mathcal{NA} \times \mathcal{NA}$  by:
  - $N \leq N$  for all nested attributes  $N \in \mathcal{NA}$ ,
  - $\lambda \leq A$  for all flat attributes  $A \in \mathcal{U}$ ,
  - $\lambda \leq N$  for all list-valued attributes  $N \in \mathcal{NA}$ ,
  - $L(N_1, \dots, N_k) \leq L(M_1, \dots, M_k)$ , if  $N_i \leq M_i$  for all  $i = 1, \dots, k$ ,
  - $L[N] \leq L[M]$ , if  $N \leq M$
- $\text{Factor}(\lambda, \text{Prime}[\text{Number}], \text{Exp}[\lambda])$  is subattribute of  $\text{Factor}(\text{Integer}, \text{Prime}[\text{Number}], \text{Exp}[\text{Number}])$
- subattribute relation  $\leq$  on nested attributes is partial order

## 2.4 Semantics on Subattributes: Projection Function

- for  $M \leq N$  define  $\pi_M^N : Dom(N) \rightarrow Dom(M)$  by:
  - $\pi_N^N : v \mapsto v$ ,
  - $\pi_\lambda^N : v \mapsto ok$ ,
  - $\pi_{L(M_1, \dots, M_k)}^{L(N_1, \dots, N_k)} : (v_1, \dots, v_k) \mapsto (\pi_{M_1}^{N_1}(v_1), \dots, \pi_{M_k}^{N_k}(v_k))$ ,
  - $\pi_{L[M']}^{L[N']} : [v_1, \dots, v_n] \mapsto [\pi_{M'}^{N'}(v_1), \dots, \pi_{M'}^{N'}(v_n)]$
- $[ ]$  mapped to itself, except when projected on  $\lambda$
- projection of  $(12, [2, 3], [2, 1])$  on  $Factor(\lambda, Prime[Number], Exp[\lambda])$  is  $(ok, [2, 3], [ok, ok])$

## 2.5 Operations on Subattributes

- $Sub(N) = \{X \in \mathcal{NA} \mid X \leq N\}$ :  $\lambda_N, Y \sqcup_N Z, Y \sqcap_N Z, Y \dot{-}_N Z$ :
  - $\lambda_N = L(\lambda_{N_1}, \dots, \lambda_{N_k})$ , if  $N = L(N_1, \dots, N_k)$ , and  $\lambda_N = \lambda$  else,
  - $X \leq Y$ :  $X \sqcup_N Y = Y$ ,  $X \sqcap_N Y = X$  and  $X \dot{-}_N Y = \lambda_N$ ,
  - $X \dot{-}_N \lambda_N = X$ ,
  - $N = L(N_1, \dots, N_k), X = L(X_1, \dots, X_k), Y = L(Y_1, \dots, Y_k)$ :  

$$X \circ_N Y = L(X_1 \circ_{N_1} Y_1, \dots, X_k \circ_{N_k} Y_k)$$
 for  $\circ \in \{\sqcup, \sqcap, \dot{-}\}$
  - $N = L[M], X = L[X'], Y = L[Y'], \circ \in \{\sqcup, \sqcap\}$ :  

$$X \circ_N Y = L[X' \circ_M Y'], \quad X \not\leq Y : X \dot{-}_N Y = L[X' \dot{-}_M Y']$$

## 2.6 The Brouwerian Algebra of Subattributes

- $(Sub(N), \leq, \sqcup_N, \sqcap_N, \dot{-}_N, N)$  is a **Brouwerian Algebra**
  - $(Sub(N), \leq, \sqcup_N, \sqcap_N)$  is a lattice
  - $N$  is top element
  - **Pseudo-Difference**  $Z \dot{-} Y$  of  $Z$  and  $Y$  in  $Sub(N)$  satisfies

$$Z \dot{-} Y \leq X \quad \text{if and only if} \quad Z \leq Y \sqcup X$$

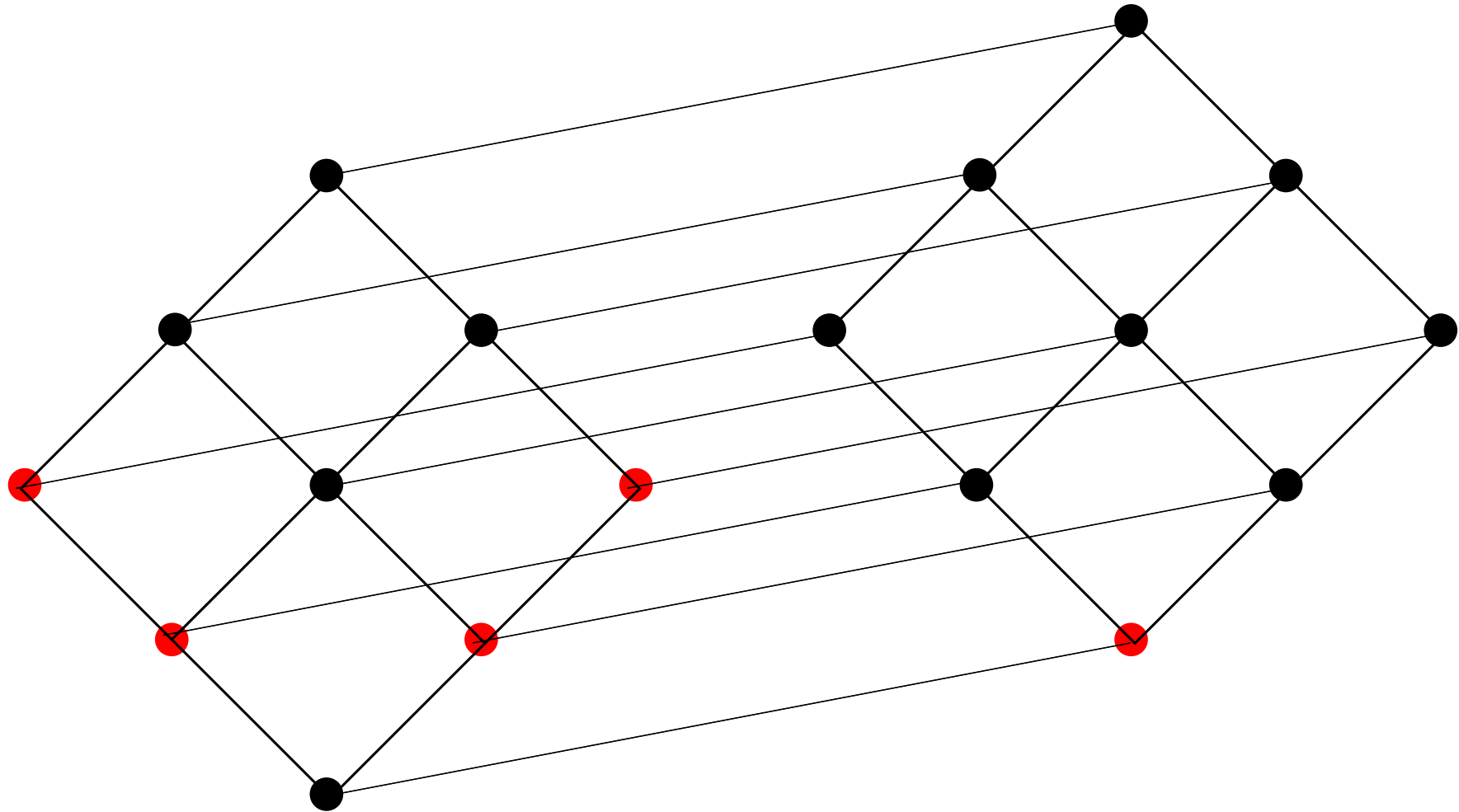
for all  $X \in Sub(N)$

- **Brouwerian Complement:**  $Y^{\mathcal{C}}_N = N \dot{-}_N Y$  satisfies

$$Y^{\mathcal{C}} \leq X \quad \text{if and only if} \quad X \sqcup Y = N$$

- bottom element  $\lambda_N = N \dot{-} N$
- $(Sub(N), \leq, \sqcup_N, \sqcap_N, (\cdot)^{\mathcal{C}}_N, \lambda_N, N)$  is not a **Boolean Algebra**
- every Brouwerian Algebra is distributive

## 2.7 Algebra of `Factor(Integer,Prime[Number],Exp[Number])`



### 3.1 Axiomatisation of Functional Dependencies

- **functional dependency** on nested attribute  $N$  is

$$X \rightarrow Y \quad \text{with} \quad X, Y \in \text{Sub}(N)$$

- finite  $r \subseteq \text{Dom}(N)$  **satisfies**  $\mathcal{X} \rightarrow \mathcal{Y}$  on  $N$  ( $\models_r \mathcal{X} \rightarrow \mathcal{Y}$ ) iff

$$\pi_X^N(t_1) = \pi_X^N(t_2) \quad \text{implies} \quad \pi_Y^N(t_1) = \pi_Y^N(t_2)$$

- The **generalized Armstrong Axioms for FDs** on  $N$

$$\frac{}{X \rightarrow Y} Y \leq X, \quad \frac{X \rightarrow Y}{X \rightarrow X \sqcup_N Y}, \quad \frac{X \rightarrow Y, Y \rightarrow Z}{X \rightarrow Z}$$

form a minimal, sound and complete set of inference rules for the implication of FDs.

## 3.2. Basis Attributes and Keys

- **subattribute basis** of  $N$ : smallest subset  $SubB(N) \subseteq Sub(N)$  st every  $X \in Sub(N)$  is  $X = \sqcup \mathcal{Z}$  for some  $\mathcal{Z} \subseteq SubB(N)$
- $MaxB(N)$  are maximal elements,  $NMaxB(N)$  the non-maximal
- subattribute basis of  $Factor(Integer, Prime[Number], Exp[Number])$ :  
 $Factor(Integer, \lambda, \lambda)$ ,  $Factor(\lambda, Prime[Number], \lambda)$ ,  $Factor(\lambda, \lambda, Exp[Number])$ ,  
 $Factor(\lambda, Prime[\lambda], \lambda)$ ,  $Factor(\lambda, \lambda, Exp[\lambda])$
- $X \in Sub(N)$  **superkey** for  $N$  wrt  $\Sigma$  iff  $\Sigma \models X \rightarrow N$
- $X \in Sub(N)$  **minimal key** for  $N$  wrt  $\Sigma$  iff  $X$  superkey and there is no proper subattribute of  $X$  also being superkey for  $N$
- $Factor(Integer, \lambda, \lambda)$  and  $Factor(\lambda, Prime[Number], Exp[Number])$  are minimal keys

### 3.3 Trivial FDs and BCNF for Lists

- $\models_r X \rightarrow Y$  for all  $r \subseteq \text{Dom}(N)$  iff  $Y \leq X$
- **trivial FD** on  $\text{Factor}(\text{Integer}, \text{Prime}[\text{Number}], \text{Exp}[\text{Number}])$ :
  - $\text{Factor}(\text{Prime}[\text{Number}]) \rightarrow \text{Factor}(\text{Prime}[\lambda])$
- $(N, \Sigma)$  in **BCNF** iff every  $X \rightarrow Y \in \Sigma^*$  trivial or  $X$  superkey
- $\text{Factor}(\text{Integer}, \text{Prime}[\text{Number}], \text{Exp}[\text{Number}])$  not in BCNF wrt  $\Sigma$  as  $\text{Factor}(\text{Prime}[\lambda]) \rightarrow \text{Factor}(\text{Exp}[\lambda])$  is not trivial and  $\text{Factor}(\text{Prime}[\lambda])$  is not a superkey

## 4.1 Redundancy and Inevitable FDs

- $(N, \Sigma)$  redundant iff  $\exists r \subseteq \text{Dom}(N)$ .  $\models_r \Sigma$  and  $\exists t_1, t_2 \in r. t_1 \neq t_2$  and  $\pi_{X \sqcup Y}^N(t_1) = \pi_{X \sqcup Y}^N(t_2)$  for some non-trivial  $X \rightarrow Y \in \Sigma$
- $\text{Factor}(\text{Prime}[\lambda]) \rightarrow \text{Factor}(\text{Exp}[\lambda])$  causes redundancies, i.e., defined in terms of non-maximal basis attribute
- non-maximal attributes cannot be separated from superattributes
- calls for better notion of redundancy
- $\Sigma_{\text{inev}} = \{X \rightarrow Y \in \Sigma^+ \mid Y \leq X \text{ or } Y \in \text{NMaxB}(N)\}$
- $\Sigma_{\text{inev}}^+$  set of **inevitable FDs** on  $N$  wrt  $\Sigma$
- $(N, \Sigma)$  **redundant** iff  $\exists r \subseteq \text{Dom}(N)$ .  $\models_r \Sigma$  and  $\exists t_1, t_2 \in r. t_1 \neq t_2$  and  $\pi_{X \sqcup Y}^N(t_1) = \pi_{X \sqcup Y}^N(t_2)$  for some  $X \rightarrow Y \in \Sigma$  not inevitable
- **Theorem:**  $(N, \Sigma)$  redundant iff  $(N, \Sigma^*)$  redundant

## 4.2 The Nested List Normal Form

- $(N, \Sigma)$  in **NLNF** iff every  $X \rightarrow Y \in \Sigma^*$  is inevitable or  $X$  superkey
- $\text{Factor}(\text{Integer}, \text{Prime}[\text{Number}], \text{Exp}[\text{Number}])$  is in NLNF wrt  $\Sigma$
- BCNF implies NLNF, but not vice versa
- **Theorem:**  $(N, \Sigma)$  in NLNF iff  $(N, \Sigma)$  non-redundant
- **Theorem:**  $(N, \Sigma)$  in NLNF iff every  $X \rightarrow Y \in \Sigma$  is inevitable or  $X$  superkey
- **Theorem:**  $(N, \Sigma)$  in NLNF iff every  $r \subseteq \text{Dom}(N)$  with  $\models_r \Sigma_{\text{key}} \cup \Sigma_{\text{inev}}^+$  satisfies  $\models_r \Sigma$

## 5.1 Update Anomalies I

- $(N, \Sigma)$  has **strong insertion anomaly** iff  $\exists r \subseteq \text{Dom}(N)$  with  $\models_r \Sigma$  and  $\exists t \notin r$  with  $\models_{r \cup \{t\}} \Sigma_{\text{key}} \cup \Sigma_{\text{inev}}^+$ , but  $\not\models_{r \cup \{t\}} \Sigma$
- **Theorem:**  $(N, \Sigma)$  in NLNF iff  $(N, \Sigma)$  does not have any strong insertion anomaly
- $(N, \Sigma)$  has **strong type-1 replacement anomaly** iff  $\exists r \subseteq \text{Dom}(N)$  with  $\models_r \Sigma$  and  $\exists t \in r, t' \in \text{Dom}(N)$  with  $\pi_K^N(t) = \pi_K^N(t')$  for minimal key  $K$  on  $N$  and  $\models_{r - \{t\} \cup \{t'\}} \Sigma_{\text{key}} \cup \Sigma_{\text{inev}}^+$  and  $\not\models_{r - \{t\} \cup \{t'\}} \Sigma$
- **Theorem:**  $(N, \Sigma)$  in NLNF iff  $(N, \Sigma)$  does not have any strong type-1 replacement anomaly

## 5.2 Update Anomalies II

- $(N, \Sigma)$  has **strong type-2 replacement anomaly** iff  $\exists r \subseteq \text{Dom}(N)$  with  $\models_r \Sigma$  and  $\exists t \in r, t' \in \text{Dom}(N)$  with  $\pi_K^N(t) = \pi_K^N(t')$  for distinguished minimal key  $K$  on  $N$  and  $\models_{r - \{t\} \cup \{t'\}} \Sigma_{\text{key}} \cup \Sigma_{\text{inev}}^+$  and  $\not\models_{r - \{t\} \cup \{t'\}} \Sigma$
- **Theorem:**  $(N, \Sigma)$  in NLNF iff  $(N, \Sigma)$  does not have any strong type-2 replacement anomaly
- $(N, \Sigma)$  has **strong type-3 replacement anomaly** iff  $\exists r \subseteq \text{Dom}(N)$  with  $\models_r \Sigma$  and  $\exists t \in r, t' \in \text{Dom}(N)$  with  $\pi_K^N(t) = \pi_K^N(t')$  for all minimal keys  $K$  on  $N$  and  $\models_{r - \{t\} \cup \{t'\}} \Sigma_{\text{key}} \cup \Sigma_{\text{inev}}^+$  and  $\not\models_{r - \{t\} \cup \{t'\}} \Sigma$
- **Theorem:**  $(N, \Sigma)$  in NLNF does not have any strong type-3 replacement anomaly

## 6.1 Extensions: More Results

- **Record and List Type:**
  - FD implication decidable in linear time
  - MVDs: minimal axiomatisation and finite implication problem
  - FDs/MVDs: minimality and finite implication problem
- **Record, List, Set and Multiset Type:**
  - sophisticated axiomatisation for FDs
  - FD implication decidable in  $\mathcal{O}(n^3 \cdot s \cdot \min\{s, n\})$
  - CVNF proposed and justified
- **XML:** several classes of FDs, Axiomatisations

## 6.2 Extensions: Future Research

- decomposition/synthesis for NLNF, CVNF
- unions, references
- interactions
- different classes of dependencies: inclusion and join dependencies
- XML: FIP and Normalisation